Mining the History Sections of Wikipedia Articles

Wolfgang Kircheis

supervised by Martin Potthast

Leipzig University

Department of Computer Science

Text Mining and Retrieval

temir.org

Motivation

Science Debates - Science Feuds - Science Wars

Illustrations by Denise Nestor

SCIENCE

THE NASTIEST FEUD IN SCIENCE

A Princeton geologist has endured decades of ridicule for arguing that the fifth extinction was caused not by an asteroid but by a series of colossal volcanic eruptions. But she's reopened that debate.

By Bianca Bosker

[*The Atlantic* 2018]

Motivation

Science Debates - Science Feuds - Science Wars

Titanic clash over CRISPR patents turns ugly

Heidi Ledford

Nature **537**, 460–461 (2016)

476 Accesses | **3** Citations | **514** Altmetric | Metrics

Accusations of impropriety feature in escalating dispute.

[*Nature 2016*]

The latest round in the CRISPR patent battle has an apparent victor, but the fight continues

Broad Institute appears to gain an advantage over the University of California and its partners who have claimed to have invented the genome editor first

11 SEP 2020 · BY JON COHEN

[*Science 2020*]

The CRISPR wars

Philip Ball

Published: April 10, 2021 • DOI: https://doi.org/10.1016/S0140-6736(21)00774-1 •

[*The Lancet 2021*]

(*Clustered Regularly Interspaced Short Palindromic Repeats)

Motivation

Innovation Scientists & Accounts

- Innovation Scientist
  - Interested in development of (scientific) innovation
  - Researches representation and discusses field-internal debate:

Motivation

Innovation Scientists & Accounts

- Innovation Scientist
  - Interested in development of (scientific) innovation
  - Researches representation and discusses field-internal debate:

Players	researchers, institutions, companies
Disputes	Who came first? Who is making which claims?
Objectivity	empirical assessment, analyse **accounts**

Motivation

Innovation Scientists & Accounts

- Innovation Scientist
  - Interested in development of (scientific) innovation
  - Researches representation and discusses field-internal debate:

Players	researchers, institutions, companies
Disputes	Who came first? Who is making which claims?
Objectivity	empirical assessment, analyse **accounts**

- Account
  - Bibliometrics, field delineation
  - Claim of relevance of scientific literature and defintion of field

Motivation

Innovation Scientists & Accounts

- Innovation Scientist
  - Interested in development of (scientific) innovation
  - Researches representation and discusses field-internal debate:

Players	researchers, institutions, companies
Disputes	Who came first? Who is making which claims?
Objectivity	empirical assessment, analyse **accounts**

- Account
  - Bibliometrics, field delineation
  - Claim of relevance of scientific literature and defintion of field

Reference Accounts	Textual Accounts
query Web of Science for 'crispr'	*The Heroes of CRISPR* [Lander 2016]
search Google Scholar for 'crispr history' 'crispr development' and 'crispr discovery'	➜ Wikipedia's history sections [Nix 2010] [Borra 2015]

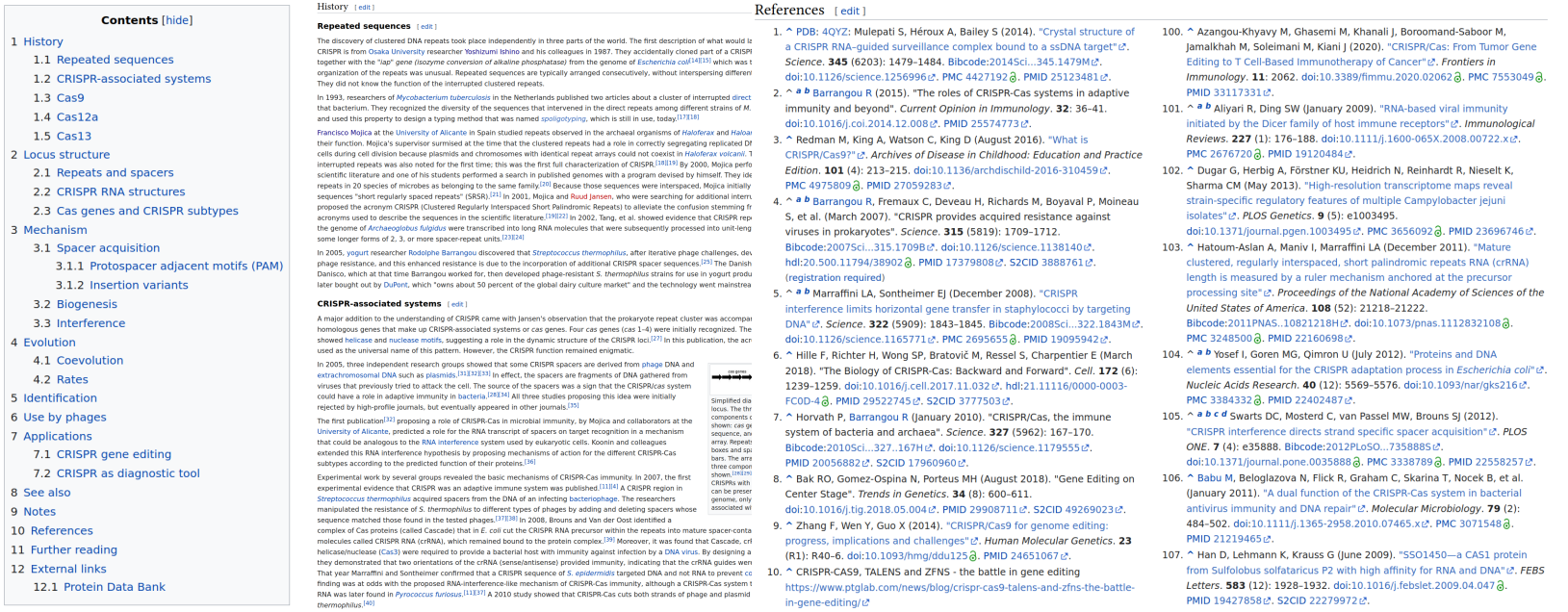
Motivation

Why Wikipedia?

- Size
- Accessibility
- Debates
- Languages
- Revisions

Background

Structuring Wikipedia Articles



Contents [hide]

History [edit]

Repeated sequences [edit]

The discovery of clustered DNA repeats took place independently in three parts of the world. The first description of what would le CRISPR is from Osaka University researcher Yoshizumi Ishino and his colleagues in 1987. They accidentally cloned part of a CRISPI together with the "*iap*" *gene (isozyme conversion of alkaline phosphatase)* from the genome of *Escherichia coli*[14][15] which was t organization of the repeats was unusual. Repeated sequences are typically arranged consecutively, without interspersing differen They did not know the function of the interrupted clustered repeats.

In 1993, researchers of *Mycobacterium tuberculosis* in the Netherlands published two articles about a cluster of interrupted direct that bacterium. They recognized the diversity of the sequences that intervened in the direct repeats among different strains of *M.* and used this property to design a typing method that was named *spoligotyping*, which is still in use, today.[17][18]

Francisco Mojica at the University of Alicante in Spain studied repeats observed in the archaeal organisms of *Haloferax* and *Haloar* their function. Mojica's supervisor surmised at the time that the clustered repeats had a role in correctly segregating replicated DI cells during cell division because plasmids and chromosomes with identical repeat arrays could not coexist in *Haloferax volcanii*. T interrupted repeats was also noted for the first time; this was the first full characterization of CRISPR.[19][19] By 2000, Mojica perfo scientific literature and one of his students performed a search in published genomes with a program devised by himself. They ide repeats in 20 species of microbes as belonging to the same family.[20] Because those sequences were interspaced, Mojica initially sequences "short regularly spaced repeats" (SRSR).[21] In 2001, Mojica and Ruud Jansen, who were searching for additional intern proposed the acronym CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) to alleviate the confusion stemming fr acronyms used to describe the sequences in the scientific literature.[19][22] In 2002, Tang, et al. showed evidence that CRISPR repe the genome of *Archaeoglobus fulgidus* were transcribed into long RNA molecules that were subsequently processed into unit-leng some longer forms of 2, 3, or more spacer-repeat units.[23][24]

In 2005, yogurt researcher Rodolphe Barrangou discovered that *Streptococcus thermophilus*, after iterative phage challenges, dev phage resistance, and this enhanced resistance is due to the incorporation of additional CRISPR spacer sequences.[25] The Danish Danisco, which at that time Barrangou worked for, then developed phage-resistant *S. thermophilus* strains for use in yogurt produ later bought out by DuPont, which "owns about 50 percent of the global dairy culture market" and the technology went mainstrea

CRISPR-associated systems [edit]

A major addition to the understanding of CRISPR came with Jansen's observation that the prokaryote repeat cluster was accompar homologous genes that make up CRISPR-associated systems or *cas* genes. Four *cas* genes (cas 1–4) were initially recognized. The showed helicase and nuclease motifs, suggesting a role in the dynamic structure of the CRISPR loci.[27] In this publication, the acr used as the universal name of this pattern. However, the CRISPR function remained enigmatic.

In 2005, three independent research groups showed that some CRISPR spacers are derived from phage DNA and extrachromosomal DNA such as plasmids.[31][32][33] In effect, the spacers are fragments of DNA gathered from viruses that previously tried to attack the cell. The source of the spacers was a sign that the CRISPR/*cas* system could have a role in adaptive immunity in bacteria.[28][34] All three studies proposing this idea were initially rejected by high-profile journals, but eventually appeared in other journals.[35]

Simplified dia locus. The thr components shown: cas ge sequence, an array. Repeat boxes and spa bars. The arra three compon shown.[28][29] CRISPRs with can be preser genome, only associated wi

The first publication[32] proposing a role of CRISPR-Cas in microbial immunity, by Mojica and collaborators at the University of Alicante, predicted a role for the RNA transcript of spacers on target recognition in a mechanism that could be analogous to the RNA interference system used by eukaryotic cells. Koonin and colleagues extended this RNA interference hypothesis by proposing mechanisms of action for the different CRISPR-Cas subtypes according to the predicted function of their proteins.[36]

Experimental work by several groups revealed the basic mechanisms of CRISPR-Cas immunity. In 2007, the first experimental evidence that CRISPR was an adaptive immune system was published.[11][8] A CRISPR region in *Streptococcus thermophilus* acquired spacers from the DNA of an infecting bacteriophage. The researchers manipulated the resistance of *S. thermophilus* to different types of phages by adding and deleting spacers whose sequence matched those found in the tested phages.[37][38] In 2008, Brouns and Van der Oost identified a complex of Cas proteins (called Cascade) that in *E. coli* cut the CRISPR RNA precursor within the repeats into mature spacer-conta molecules called CRISPR RNA (crRNA), which remained bound to the protein complex.[39] Moreover, it was found that Cascade, crR helicase/nuclease (Cas3) were required to provide a bacterial host with immunity against infection by a DNA virus. By designing a they demonstrated that two orientations of the crRNA (sense/antisense) provided immunity, indicating that the crRNA guides wea That year Marraffini and Sontheimer confirmed that a CRISPR sequence of *S. epidermidis* targeted DNA and not RNA to prevent co finding was at odds with the proposed RNA-interference-like mechanism of CRISPR-Cas immunity, although a CRISPR-Cas system t RNA was later found in *Pyrococcus furiosus*.[11][37] A 2010 study showed that CRISPR-Cas cuts both strands of phage and plasmid *thermophilus*.[40]

References [edit]

1. ^ PDB: 4QYZ: Mulepati S, Héroux A, Bailey S (2014). "Crystal structure of a CRISPR RNA–guided surveillance complex bound to a ssDNA target". *Science*. **345** (6203): 1479–1484. Bibcode:2014Sci...345.1479M. doi:10.1126/science.1256996. PMC 4427192. PMID 25123481.
2. ^ **a b** Barrangou R (2015). "The roles of CRISPR-Cas systems in adaptive immunity and beyond". *Current Opinion in Immunology*. **32**: 36–41. doi:10.1016/j.coi.2014.12.008. PMID 25574773.
3. ^ Redman M, King A, Watson C, King D (August 2016). "What is CRISPR/Cas9?". *Archives of Disease in Childhood: Education and Practice Edition*. **101** (4): 213–215. doi:10.1136/archdischild-2016-310459. PMC 4975809. PMID 27059283.
4. ^ **a b** Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, et al. (March 2007). "CRISPR provides acquired resistance against viruses in prokaryotes". *Science*. **315** (5819): 1709–1712. Bibcode:2007Sci...315.1709B. doi:10.1126/science.1138140. hdl:20.500.11794/38902. PMID 17379808. S2CID 3888761. (registration required)
5. ^ **a b** Marraffini LA, Sontheimer EJ (December 2008). "CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA". *Science*. **322** (5909): 1843–1845. Bibcode:2008Sci...322.1843M. doi:10.1126/science.1165771. PMC 2695655. PMID 19095942.
6. ^ Hille F, Richter H, Wong SP, Bratovič M, Ressel S, Charpentier E (March 2018). "The Biology of CRISPR-Cas: Backward and Forward". *Cell*. **172** (6): 1239–1259. doi:10.1016/j.cell.2017.11.032. hdl:21.11116/0000-0003-FC0D-4. PMID 29522745. S2CID 3777503.
7. ^ Horvath P, Barrangou R (January 2010). "CRISPR/Cas, the immune system of bacteria and archaea". *Science*. **327** (5962): 167–170. Bibcode:2010Sci...327..167H. doi:10.1126/science.1179555. PMID 20056882. S2CID 17960960.
8. ^ Bak RO, Gomez-Ospina N, Porteus MH (August 2018). "Gene Editing on Center Stage". *Trends in Genetics*. **34** (8): 600–611. doi:10.1016/j.tig.2018.05.004. PMID 29908711. S2CID 49269023.
9. ^ Zhang F, Wen Y, Guo X (2014). "CRISPR/Cas9 for genome editing: progress, implications and challenges". *Human Molecular Genetics*. **23** (R1): R40–6. doi:10.1093/hmg/ddu125. PMID 24651067.
10. ^ CRISPR-CAS9, TALENS and ZFNS - the battle in gene editing https://www.ptglab.com/news/blog/crispr-cas9-talens-and-zfns-the-battle-in-gene-editing/
100. ^ Azangou-Khyavy M, Ghasemi M, Khanali J, Boroomand-Saboor M, Jamalkhah M, Soleimani M, Kiani J (2020). "CRISPR/Cas: From Tumor Gene Editing to T Cell-Based Immunotherapy of Cancer". *Frontiers in Immunology*. **11**: 2062. doi:10.3389/fimmu.2020.02062. PMC 7553049. PMID 33117331.
101. ^ **a b** Aliyari R, Ding SW (January 2009). "RNA-based viral immunity initiated by the Dicer family of host immune receptors". *Immunological Reviews*. **227** (1): 176–188. doi:10.1111/j.1600-065X.2008.00722.x. PMC 2676720. PMID 19120484.
102. ^ Dugar G, Herbig A, Förstner KU, Heidrich N, Reinhardt R, Nieselt K, Sharma CM (May 2013). "High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates". *PLOS Genetics*. **9** (5): e1003495. doi:10.1371/journal.pgen.1003495. PMC 3656092. PMID 23696746.
103. ^ Hatoum-Aslan A, Maniv I, Marraffini LA (December 2011). "Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site". *Proceedings of the National Academy of Sciences of the United States of America*. **108** (52): 21218–21222. Bibcode:2011PNAS..10821218H. doi:10.1073/pnas.1112832108. PMC 3248500. PMID 22160698.
104. ^ **a b** Yosef I, Goren MG, Qimron U (July 2012). "Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*". *Nucleic Acids Research*. **40** (12): 5569–5576. doi:10.1093/nar/gks216. PMC 3384332. PMID 22402487.
105. ^ **a b c d** Swarts DC, Mosterd C, van Passel MW, Brouns SJ (2012). "CRISPR interference directs strand specific spacer acquisition". *PLOS ONE*. **7** (4): e35888. Bibcode:2012PLoSO...735888S. doi:10.1371/journal.pone.0035888. PMC 3338789. PMID 22558257.
106. ^ Babu M, Beloglazova N, Flick R, Graham C, Skarina T, Nocek B, et al. (January 2011). "A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair". *Molecular Microbiology*. **79** (2): 484–502. doi:10.1111/j.1365-2958.2010.07465.x. PMC 3071548. PMID 21219465.
107. ^ Han D, Lehmann K, Krauss G (June 2009). "SSO1450—a CAS1 protein from Sulfolobus solfataricus P2 with high affinity for RNA and DNA". *FEBS Letters*. **583** (12): 1928–1932. doi:10.1016/j.febslet.2009.04.047. PMID 19427858. S2CID 22279972.

- Sections, subsections and subsubsections
- Embedded elements like diagrams, charts, tables
- *References* and *Further Reading*

Background

Structuring Wikipedia

- Article, Talk, Revision history, Media, ...

  CRISPR

  [revid 583140353, 20:47, 24 November 2013] ➜ [revid 583141384, 20:55, 24 November 2013]

  CRISPR gene editing

  [revid 942549511, 10:30, 25 February 2020] ➜ [revid 942550392, 10:39, 25 February 2020]

Background

Structuring Wikipedia

- Article, Talk, Revision history, Media, ...

  CRISPR

  [revid 583140353, 20:47, 24 November 2013] ➜ [revid 583141384, 20:55, 24 November 2013]

  CRISPR gene editing

  [revid 942549511, 10:30, 25 February 2020] ➜ [revid 942550392, 10:39, 25 February 2020]

- Wikipedia API:

  `https://en.wikipedia.org/w/api.php?format=json&action=query&titles=CRISPR&prop=revisions&rvlimit=1&rvdir=newer&rvslots=*&rvstartid=373434559&rvprop=comment|content|contentmodel|flagged|flags|ids|oresscores|parsedcomment|roles|sha1|size|slotsha1|slotsize|tags|timestamp|user|userid`

Background

Structuring Wikipedia

- Article, Talk, Revision history, Media, ...

  CRISPR

  [revid 583140353, 20:47, 24 November 2013] ➜ [revid 583141384, 20:55, 24 November 2013]

  CRISPR gene editing

  [revid 942549511, 10:30, 25 February 2020] ➜ [revid 942550392, 10:39, 25 February 2020]

- Wikipedia API:

  `https://en.wikipedia.org/w/api.php?format=json&action=query&titles=CRISPR&prop=revisions&rvlimit=1&rvdir=newer&rvslots=*&rvstartid=373434559&rvprop=comment|content|contentmodel|flagged|flags|ids|oresscores|parsedcomment|roles|sha1|size|slotsha1|slotsize|tags|timestamp|user|userid`

- Wikimedia dumps:

  `https://dumps.wikimedia.org/enwiki/20211101`

Background

Structuring Wikipedia

- Article, Talk, Revision history, Media, ...

  CRISPR

  [revid 583140353, 20:47, 24 November 2013] ➜ [revid 583141384, 20:55, 24 November 2013]

  CRISPR gene editing

  [revid 942549511, 10:30, 25 February 2020] ➜ [revid 942550392, 10:39, 25 February 2020]

- Wikipedia API:

  `https://en.wikipedia.org/w/api.php?format=json&action=query&titles=CRISPR&prop=revisions&rvlimit=1&rvdir=newer&rvslots=*&rvstartid=373434559&rvprop=comment|content|contentmodel|flagged|flags|ids|oresscores|parsedcomment|roles|sha1|size|slotsha1|slotsize|tags|timestamp|user|userid`

- Wikimedia dumps:

  `https://dumps.wikimedia.org/enwiki/20211101`

- Issues: redirects, missing links, mismatches API and dump, **Wikitext**

Background

Wikitext vs HTML

- Wikitext
  - Readily available: standard format in Wikipedia REST API and Wikimedia dumps
  - Formatting: incomplete, in-line references, artifacts [Mitrevski 2020]

```
===Discovery and properties===

The existence of the DNA fragments, which are known as CRISPR today, was discovered in 1987 in the bacterium [[Escherichia coli|E. coli]].<ref
name=\"pmid3316184\">{{cite journal |author=Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A |title=Nucleotide sequence of the iap gene,
responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. |journal=J Bacteriol |volume=169
|issue=12 |pages=5429-33 |year=1987 |pmid=3316184}}</ref> In 2002 it was announced that there exist similar structures in the genome of many
different prokaryotes, and the name CRISPR was coined CRISPR.<ref name=\"pmid11952905\">{{cite journal |author=Jansen R, Embden JD, Gaastra W,
Schouls LM |title=Identification of genes that are associated with DNA repeats in prokaryotes. |journal=Mol Microbiol |volume=43 |issue=6 |
pages=1565-75 |year=2002 |pmid=11952905}}</ref> In addition, a group of genes discovered in all investigated organisms near the locus of CRISPR
repeats called CAS (CRISPR-associated) genes were were identified.<ref name=\"pmid11952905\"></ref>
```

Background

Wikitext vs HTML

- Wikitext
  - Readily available: standard format in Wikipedia REST API and Wikimedia dumps
  - Formatting: incomplete, in-line references, artifacts [Mitrevski 2020]

```
===Discovery and properties===

The existence of the DNA fragments, which are known as CRISPR today, was discovered in 1987 in the bacterium [[Escherichia coli|E. coli]].<ref
name=\"pmid3316184\">{{cite journal |author=Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A |title=Nucleotide sequence of the iap gene,
responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. |journal=J Bacteriol |volume=169
|issue=12 |pages=5429-33 |year=1987 |pmid=3316184}}</ref> In 2002 it was announced that there exist similar structures in the genome of many
different prokaryotes, and the name CRISPR was coined CRISPR.<ref name=\"pmid11952905\">{{cite journal |author=Jansen R, Embden JD, Gaastra W,
Schouls LM |title=Identification of genes that are associated with DNA repeats in prokaryotes. |journal=Mol Microbiol |volume=43 |issue=6 |
pages=1565-75 |year=2002 |pmid=11952905}}</ref> In addition, a group of genes discovered in all investigated organisms near the locus of CRISPR
repeats called CAS (CRISPR-associated) genes were were identified.<ref name=\"pmid11952905\"></ref>
```

- HTML
  - Information: HTML more extensive than Wikitext
  - Structure: DOM tree
  - Not readily available: requires parsing Wikitext or scraping HTML

```
<h3><span class="mw-headline" id="Discovery_and_properties">Discovery and properties</span></h3>

<p>The existence of the DNA fragments, which are known as CRISPR today, was discovered in 1987 in the bacterium <a href="/wiki/Escherichia_coli"
title="Escherichia coli">E. coli</a>.<sup id="cite_ref-pmid3316184_2-0" class="reference"><a href="#cite_note-pmid3316184-2">&#91;2&#93;</a></sup> In
2002 it was announced that there exist similar structures in the genome of many different prokaryotes, and the name CRISPR was coined CRISPR.<sup
id="cite_ref-pmid11952905_3-0" class="reference"><a href="#cite_note-pmid11952905-3">&#91;3&#93;</a></sup> In addition, a group of genes discovered
in all investigated organisms near the locus of CRISPR repeats called CAS (CRISPR-associated) genes were were identified.<sup id="cite_ref-
pmid11952905_3-1" class="reference"><a href="#cite_note-pmid11952905-3">&#91;3&#93;</a></sup></p>
```

Background

Wikitext vs HTML

- Wikitext
  - Readily available: standard format in Wikipedia REST API and Wikimedia dumps
  - Formatting: incomplete, in-line references, artifacts [Mitrevski 2020]

```
===Discovery and properties===

The existence of the DNA fragments, which are known as CRISPR today, was discovered in 1987 in the bacterium [[Escherichia coli|E. coli]].<ref
name=\"pmid3316184\">{{cite journal |author=Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A |title=Nucleotide sequence of the iap gene,
responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. |journal=J Bacteriol |volume=169
|issue=12 |pages=5429-33 |year=1987 |pmid=3316184}}</ref> In 2002 it was announced that there exist similar structures in the genome of many
different prokaryotes, and the name CRISPR was coined CRISPR.<ref name=\"pmid11952905\">{{cite journal |author=Jansen R, Embden JD, Gaastra W,
Schouls LM |title=Identification of genes that are associated with DNA repeats in prokaryotes. |journal=Mol Microbiol |volume=43 |issue=6 |
pages=1565-75 |year=2002 |pmid=11952905}}</ref> In addition, a group of genes discovered in all investigated organisms near the locus of CRISPR
repeats called CAS (CRISPR-associated) genes were were identified.<ref name=\"pmid11952905\"></ref>
```

- HTML
  - Information: HTML more extensive than Wikitext
  - Structure: DOM tree
  - Not readily available: requires parsing Wikitext or scraping HTML

```
<h3><span class="mw-headline" id="Discovery_and_properties">Discovery and properties</span></h3>

<p>The existence of the DNA fragments, which are known as CRISPR today, was discovered in 1987 in the bacterium <a href="/wiki/Escherichia_coli"
title="Escherichia coli">E. coli</a>.<sup id="cite_ref-pmid3316184_2-0" class="reference"><a href="#cite_note-pmid3316184-2">&#91;2&#93;</a></sup> In
2002 it was announced that there exist similar structures in the genome of many different prokaryotes, and the name CRISPR was coined CRISPR.<sup
id="cite_ref-pmid11952905_3-0" class="reference"><a href="#cite_note-pmid11952905-3">&#91;3&#93;</a></sup> In addition, a group of genes discovered
in all investigated organisms near the locus of CRISPR repeats called CAS (CRISPR-associated) genes were were identified.<sup id="cite_ref-
pmid11952905_3-1" class="reference"><a href="#cite_note-pmid11952905-3">&#91;3&#93;</a></sup></p>
```

- WikiHist.html [Mitrevski 2020]
  - 580M HTML revisions of 5.8M articles (7 TB), 1 January 2001 to 1 March 2019

Challenges

- Lack of structure and consistency
  - no consistent templates, plethora of page types [Aprosio 2015]
  - Random sample of 100 persons' pages:
    20 with section called *Biography* or *Life* vs 80 with person's biography in sections
  - `https://en.wikipedia.org/wiki/Leonard_Bernstein` vs
    `https://en.wikipedia.org/wiki/Judy_Holliday`
  - article-as-concept assumption and sub-article matching problem [Lin 2017]

Challenges

- Lack of structure and consistency
  - no consistent templates, plethora of page types [Aprosio 2015]
  - Random sample of 100 persons' pages:
    20 with section called *Biography* or *Life* vs 80 with person's biography in sections
  - `https://en.wikipedia.org/wiki/Leonard_Bernstein` vs
    `https://en.wikipedia.org/wiki/Judy_Holliday`
  - article-as-concept assumption and sub-article matching problem [Lin 2017]

- Recognising history sections
  - Using section titles
    - 'history', 'development', ...
    - 'discovery', 'predecessors', ...
  - Articles lacking section subdivisions or consistent headings [Field 2020]

Challenges

- Lack of structure and consistency
  - no consistent templates, plethora of page types [Aprosio 2015]
  - Random sample of 100 persons' pages:
    20 with section called *Biography* or *Life* vs 80 with person's biography in sections
  - `https://en.wikipedia.org/wiki/Leonard_Bernstein` vs
    `https://en.wikipedia.org/wiki/Judy_Holliday`
  - article-as-concept assumption and sub-article matching problem [Lin 2017]
- Recognising history sections
  - Using section titles
    - 'history', 'development', ...
    - 'discovery', 'predecessors', ...
  - Articles lacking section subdivisions or consistent headings [Field 2020]
- Article types
  - Technology, application, person, country, ...
  - Pre-select article candidates

Solutions

- Section Tree Reader
  - Use DOM tree of HTML and headings
  - Build tree of sections, their texts and metadata
  - Wikitext-based version for (quick) comparison

Solutions

- Section Tree Reader
  - Use DOM tree of HTML and headings
  - Build tree of sections, their texts and metadata
  - Wikitext-based version for (quick) comparison

- Use sections' titles
  - Baseline approaches: verbatim matching, frequent tokens
  - CRF or SVM [Aprosio 2015]

Solutions

- Section Tree Reader
  - Use DOM tree of HTML and headings
  - Build tree of sections, their texts and metadata
  - Wikitext-based version for (quick) comparison
- Use sections' titles
  - Baseline approaches: verbatim matching, frequent tokens
  - CRF or SVM [Aprosio 2015]
- Use sections' contents
  - Baseline approaches: verbatim matching, frequent tokens
  - Thematically map contents of section: topic modeling [Liu 2016]
  - Granularity of topics
  - Title section using contents [Field 2020]
  - Feedback loop: map actual titles to generated titles

Analysis

- Case Study
  - CRISPR
  - 849 article candidate dataset
  - ~30 article set with more than 5 publications
  - 11 article set based on qualitative and quantitative assessments

Analysis

- Case Study
  - CRISPR
  - 849 article candidate dataset
  - ~30 article set with more than 5 publications
  - 11 article set based on qualitative and quantitative assessments

- Evaluation
  - Using technology articles
  - compare baseline and advances approaches
  - precision and recall

Analysis

- Case Study
  - CRISPR
  - 849 article candidate dataset
  - ~30 article set with more than 5 publications
  - 11 article set based on qualitative and quantitative assessments
- Evaluation
  - Using technology articles
  - compare baseline and advances approaches
  - precision and recall
- Typology
  - Find history sections and then check article type
  - Qualitative assessment

Mining the History Sections of Wikipedia Articles

Wolfgang Kircheis

supervised by Martin Potthast

Leipzig University

Department of Computer Science

Text Mining and Retrieval

temir.org

Recap

Innovation Scientists & Accounts

- Innovation scientist interested in development of (scientific) innovation
- Researchers, institutions, companies, claims, relevance

Recap

Innovation Scientists & Accounts

- Innovation scientist interested in development of (scientific) innovation
- Researchers, institutions, companies, claims, relevance

- Reference Accounts
  - References (Google Scholar, Semantic Scholar, Web of Science)
  - Field-delineated bibliographies
- Textual Accounts
  - Popular science magazines
  - Publications detailing development of technology
  - Wikipedia articles and their history sections

Recap

Innovation Scientists & Accounts

- Innovation scientist interested in development of (scientific) innovation
- Researchers, institutions, companies, claims, relevance

- Reference Accounts
  - References (Google Scholar, Semantic Scholar, Web of Science)
  - Field-delineated bibliographies
- Textual Accounts
  - Popular science magazines
  - Publications detailing development of technology
  - Wikipedia articles and their history sections

- Why Wikipedia?
  - Size
  - Accessibility
  - Debates
  - Languages
  - Revisions

Outline

1. Heuristics
   - Science & Technology: filter out non-relevant articles using categories
   - History Sections: find history sections using heading

2. Classification
   - Train classifier to find history section
   - Articles with designated history sections as training data [Aprosio 2015]

3. Field Study
   - Apply classifier to revision history of selected article(s)
   - Analyse revisions (works cited, researchers, ...)

Preliminary Analysis

Wikidump & Architecture

- enwiki dump from 1 June 2021 (17.4 GB)
  - Recombined articles, templates, media/file descriptions and primary meta-pages
  - 15,914,644 data entries in dump

Preliminary Analysis

Wikidump & Architecture

- enwiki dump from 1 June 2021 (17.4 GB)
  - Recombined articles, templates, media/file descriptions and primary meta-pages
  - 15,914,644 data entries in dump
- Architecture
  - WikipediaDumpReader to extract title, pageid, revid, timestamp and wikitext of articles
  - WikitextReader to extract categories and headings, build heading and section tree from Wikitext

Preliminary Analysis

Wikidump & Architecture

- enwiki dump from 1 June 2021 (17.4 GB)
  - Recombined articles, templates, media/file descriptions and primary meta-pages
  - 15,914,644 data entries in dump
- Architecture
  - WikipediaDumpReader to extract title, pageid, revid, timestamp and wikitext of articles
  - WikitextReader to extract categories and headings, build heading and section tree from Wikitext
- Results
  - 6,002,210 articles with extractable section tree
  - 2,943,189 categories
  - 692,514 articles (11.54%) with section heading 'history'

Outline

1. Heuristics

2. Classification

3. Field Study

Heuristics

Step 1.1 - 1.3 2022-01-05 18:13:14

	categories	articles	section contains 'history'	section 'history'
any	2,943,189	6,002,210	831,708 (13.86%)	692,514 (11.54%)
science $\vee$ technology	168,187	104,155	13,965 (13.41%)	11,145 (10.70%)
(science $\vee$ technology) $\wedge$ $\neg$ (person $\vee$ company)	98,004	57,681	10,308 (17.87%)	8,066 (13.98%)

Heuristics

Step 1.2 + 1.3 2022-01-05 18:13:14

- Articles with extractable sections only
- Categories:
  - 'science' or 'technology'
  - 'people' or 'companies'
- Headings: 'history' (sub-string and exact match)
- science $\vee$ technology

categories	articles	heading with 'history'	heading 'history'
168,187	104,155	13,965 (13.41%)	11,145 (10.70%)

N=104,155, n=50: 24 S&T (52%) (24 people, 2 companies)

- (science $\vee$ technology) $\wedge \neg$ (person $\vee$ company)

categories	articles	heading with 'history'	heading 'history'
98,004	57,681	10,308 (17.87%)	8,066 (13.98%)

N=57,681, n=50: 20 S&T (40%) (13 media, 7 people, 6 institutions, 4 journals)

Heuristics

Step 1.4 2022-01-05 18:13:14

- 'science', 'technolog': match in category (no 'y')
- Manually check most frequent categories
- media ('films', 'movies', 'series', 'anime', 'manga', 'books', 'novels', 'screenplays', 'fiction', 'stories', 'games'), people ('people', 'fellows', 'members', 'male', 'female', 'writers', 'authors', 'alumni'), institutions ('institution', 'companies', 'colleges', 'universities', 'council', 'convention'), other ('journals', 'magazines', 'births', 'deaths'): no match in categories

categories	articles	heading with 'history'	heading 'history'
43,612	27,819	7,340 (26.38%)	6,288 (22.60%)

N=27,819, n=50: 33 S&T (66%) (10 institutions, 4 events, 3 lists)

Heuristics

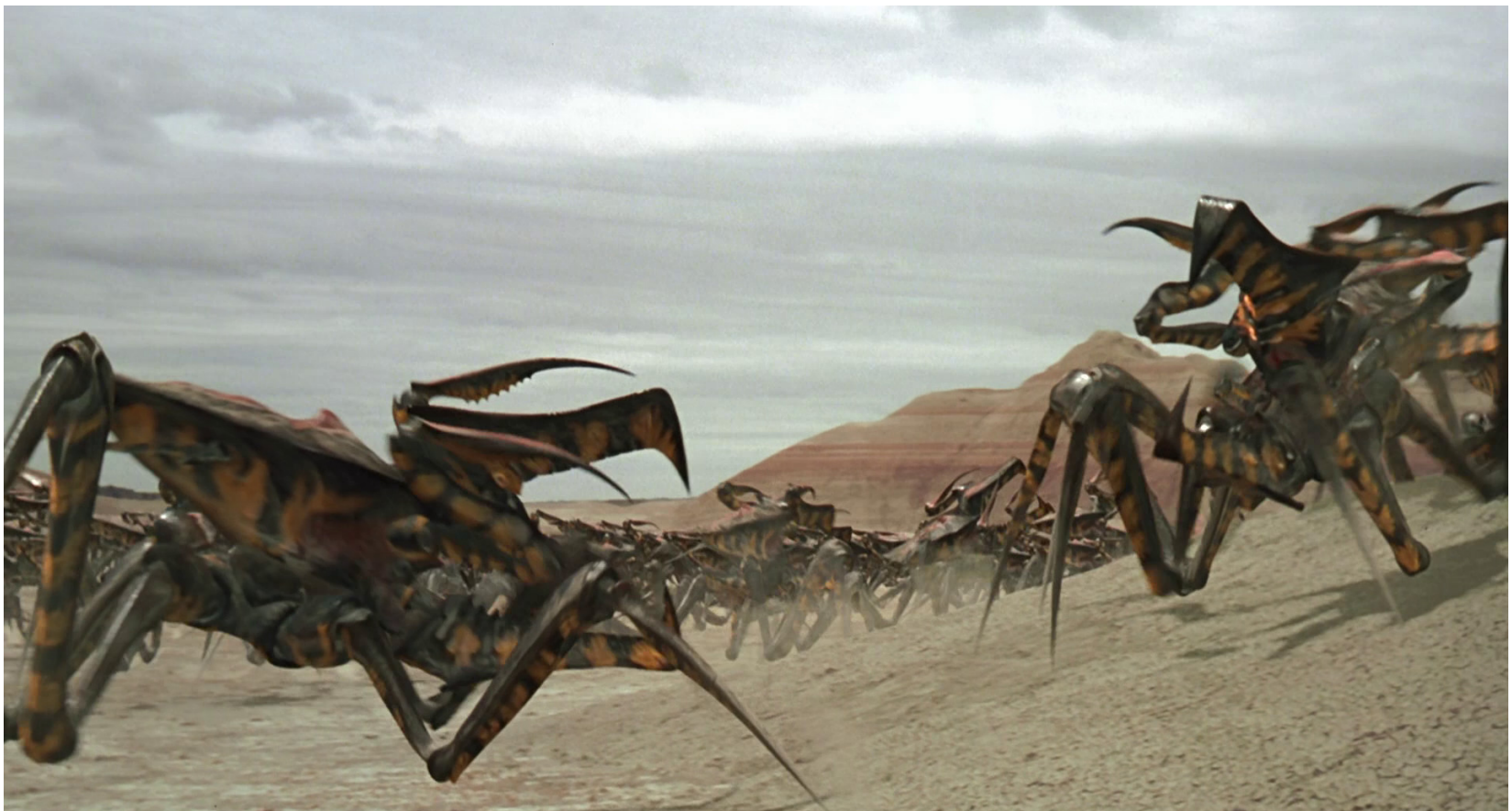
Step 2 2022-01-13 19:12:18

- 'science', 'technolog': match in category
- '^\d\d\d\ds? in', '^List of': no match in title
- Manually check most frequent categories and category tokens
- media ('films', 'movies', 'series', 'anime', 'manga', +8), people ('people', 'fellows', 'members', 'male', 'female', +3), institutions ('institution', 'companies', 'colleges', 'institute', 'department', +15), events ('events', 'festivals', 'conventions', 'awards', 'conferences'), other ('journals', 'magazines', 'births', 'deaths', 'buildings', +3): no match in categories

categories	articles	heading with 'history'	heading 'history'
18,034	17,085	4,454 (26.07%)	3,847 (22.52%)

Heuristics

BUGS!!!!!



Heuristics

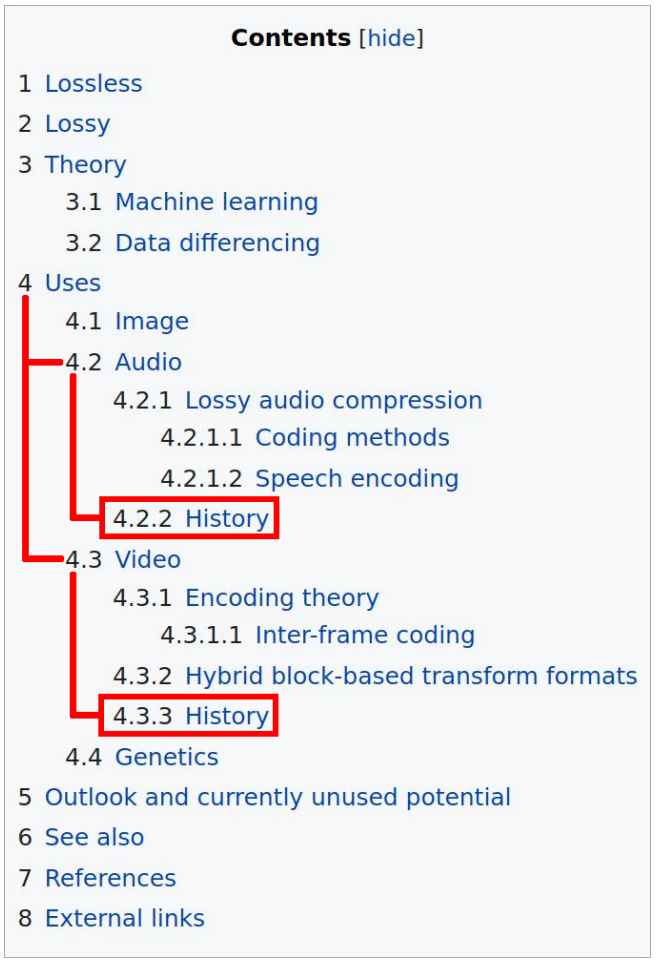
Step 3 2022-03-14 16:32:11

- 'science', 'technolog': match in category
- 'list', 'history', 'science and technology', '\d\d\d\d': no match in title
- media ('films', 'movies', 'series', 'anime', 'manga', +8), people ('people', 'fellows', 'members', 'male', 'female', +3), institutions ('institution', 'companies', 'colleges', 'institute', 'department', +16), events ('events', 'festivals', 'conventions', 'awards', 'conferences'), other ('journals', 'magazines', 'births', 'deaths', 'buildings', +3): no match in categories
- heading with 'history' → heading with 'histor[y|i]'

categories	articles	heading with 'histor[y|i]'	heading 'history'
17,840	16,961	4,743 (27.96%)	3,953 (23.31%)

Heuristics

Refactoring: Top vs Any Level



Contents [hide]

1 Lossless
2 Lossy
3 Theory
  3.1 Machine learning
  3.2 Data differencing
4 Uses
  4.1 Image
  4.2 Audio
    4.2.1 Lossy audio compression
      4.2.1.1 Coding methods
      4.2.1.2 Speech encoding
    4.2.2 History
  4.3 Video
    4.3.1 Encoding theory
      4.3.1.1 Inter-frame coding
    4.3.2 Hybrid block-based transform formats
    4.3.3 History
  4.4 Genetics
5 Outlook and currently unused potential
6 See also
7 References
8 External links

Heuristics

Step 3 2022-03-14 16:32:11

- 'science', 'technolog': match in category
- 'list', 'history', 'science and technology', '\d\d\d\d': no match in title
- media ('films', 'movies', 'series', 'anime', 'manga', +8), people ('people', 'fellows', 'members', 'male', 'female', +3), institutions ('institution', 'companies', 'colleges', 'institute', 'department', +16), events ('events', 'festivals', 'conventions', 'awards', 'conferences'), other ('journals', 'magazines', 'births', 'deaths', 'buildings', +3): no match in categories

categories	articles
17,840	16,961

	heading with 'histor[y|i]'	heading 'history'
any level	4,743 (27.96%)	3,953 (23.30%)
top level	4,564 (26.91%)	3,861 (22.76%)

Heuristics

Step 3 2022-03-14 16:32:11

- 10+ subsections at top level
  - → 'See also', 'References', 'Bibliography', 'Further reading', 'External links'
  - → articles sufficiently long
  - → exploitable structure (cf. [Aprosio 2015])

- N=1,559, n=100: 88 S&T (88%)

- High false positive science-and-technology: glossaries, indices, authorities, esoteric/pseudoscience articles among results

- **No heading with 'histor[y|i]'**
  N=813, n=50, 21 have history section (42%)
  Heading 'history' at top level
  N=746, n=50, 49 have history section (98%)

Heuristics

Step 4 2022-03-17 19:58:04

- 'science', 'technolog': match in category
- 'list', 'history', 'science and technology', '\d\d\d\d': no match in title
- media ('films', 'movies', 'series', 'albums', 'anime', +9), people ('people', 'fellows', 'members', 'male', 'female', +3), institutions ('institution', 'companies', 'colleges', 'institute', 'department', +17), events ('events', 'festivals', 'conventions', 'awards', 'conferences', +2), places ('villages', 'towns', 'cities', 'countries', 'states', +1), other ('journals', 'magazines', 'births', 'deaths', 'buildings', +4), collections ('lists', 'indixes', 'indices', 'glossaries'), not science ('esoteric', 'pseudoscience'): no match in categories

categories	articles
14,667	15,177

	heading with 'histor[y|i]'	heading 'history'
any level	4,093 (26.97%)	3,419 (22.53%)
top level	3,933 (25.91%)	3,332 (21.95%)

Heuristics

Step 4 2022-03-17 19:58:04

- 10+ subsections at top level
  - → 'See also', 'References', 'Bibliography', 'Further reading', 'External links'
  - → articles sufficiently long
  - → exploitable structure (cf. [Aprosio 2015])
- N=1,333, n=100: 96 S&T (96%)
- Better science-and-technology precision, some articles difficult to assess:
  - categories: 'by country', 'schools', 'occupations', 'parks', 'districts'
  - articles: 'Hydrogen', 'Statistics', 'Geochemistry', 'Political sociology'
- **No heading with 'histor[y|i]'**
  N=670, n=50, 19 have history section (38%)
  Heading 'history' at top level
  N=663, n=50, 49 have history section (98%)
- High false negative history section: 'Development', 'Background', 'Overview', 'Timeline', 'Origins', 'Roots', 'Milestones', 'Evolution (of...)'

Heuristics

Step 5 2022-03-29 01:39:26

- 'technolog': match in category
- 'list of', 'index of', ' in ', ' on ', 'history', 'institution', +22: no match in title
- media ('films', 'movies', 'series', 'albums', 'anime', +10), people ('people', 'fellows', 'members', 'male', 'female', +4), institutions ('institution', 'companies', 'schools', 'colleges', 'institute', +18), events ('events', 'festivals', 'conventions', 'awards', 'conferences', +2), places ('villages', 'towns', 'cities', 'countries', 'states', +4), other ('journals', 'magazines', 'births', 'deaths', 'buildings', +4), collections ('lists', 'indixes', 'indices', 'glossaries'), not science ('esoteric', 'pseudoscience'): no match in categories

categories	articles
8,752	8,402

	heading with 'histor[y|i]'	heading 'history'
any level	2,363 (28.12%)	2,068 (24.61%)
top level	2,289 (27.24%)	2,021 (24.05%)

Heuristics

Step 5 2022-03-29 01:39:26

- 3+ subsections at top level
  - → excluding 'See also', 'References', 'Bibliography', 'Further reading', 'External links'
  - → articles sufficiently long
  - → exploitable structure (cf. [Aprosio 2015])
- 9 labellers, 10 batches, 1 batch labelled by 5 labellers
- N=4,409, n=650, CL=99%, CI<5%: 621 S&T (95.53%)
- Good science-and-technology precision
- **No heading with 'histor[y|i]'**
  N=2,825, n=340, CL=95%, CI=5% 45 have history section (13.24%)
  Heading 'history' at top level
  N=1,584, n=310, CL=95%, CI=5% 307 have history section (99.03%)
- Suboptimal false negative history section

DESIGNATED HISTORY SECTION $\rightarrow$ HISTORY

NO DESIGNATED HISTORY SECTION $\not\rightarrow$ NO HISTORY

Heuristics

Inter-Labeller (Dis)agreement

- Inter-labeller agreement (Cohen's Kappa)

	labeller 02	labeller 04	labeller 08	labeller 06	labeller 10
labeller 02	-	(84.90%)	(90.75%)	(71.51%)	(81.64%)
labeller 04	(84.90%)	-	(93.94%)	(69.94%)	(85.11%)
labeller 08	(90.75%)	(93.94%)	-	(75.29%)	(90.90%)
labeller 06	(71.51%)	(69.94%)	(75.29%)	-	(75.17%)
labeller 10	(81.64%)	(85.11%)	(90.90%)	(75.17%)	-

- Inter-labeller disagreement (all articles without heading with 'histor[y|i]')

	labeller 02	labeller 04	labeller 08	labeller 06	labeller 10
Substantial equivalence	✓	×	✓	✓	✓
Gremlin	✓	?	✓	✓	✓
Sound Blaster X7	×	×	×	✓	×
Moving iron speaker	✓	×	×	×	×
IBTS Greenhouse	×	×	×	✓	×
W Motors Lykan HyperSport	×	×	×	✓	×
Forward osmosis	×	×	×	✓	×
Strategic communication	×	×	×	×	✓
Carbon nanotube field-effect transistor	×	×	×	✓	×
Chimney (locomotive)	×	×	×	✓	?
Hunveyor	×	×	×	✓	✓
Telecentre	✓	×	×	×	×
Wendelstein 7-AS	✓	×	×	✓	×

Heuristics

Top 50 Categories with Frequency

1	film and video technology	495
2	emerging technologies	342
3	biotechnology	330
4	television technology	280
5	information technology management	196
6	engine technology	195
7	display technology	173
8	nanotechnology	160
9	mobile technology	144
10	gas technologies	129
11	cooling technology	121
12	radio technology	121
13	assistive technology	118
14	automotive technologies	110
15	nuclear technology	103
16	educational technology	99
17	waste treatment technology	99
18	american inventions	98
19	sustainable technologies	97
20	drilling technology	97
21	articles containing video clips	95
22	science and technology in poland	94
23	military technology	91
24	vehicle safety technologies	90
25	microwave technology	89

26	television terminology	86
27	technology in society	86
28	molecular biology	85
29	operating system technology	84
30	vehicle technology	82
31	automotive transmission technologies	81
32	microtechnology	81
33	sound production technology	81
34	power station technology	79
35	automotive suspension technologies	77
36	appropriate technology	75
37	hydrogen technologies	73
38	automotive technology tradenames	73
39	membrane technology	71
40	packets (information technology)	66
41	broadcast engineering	64
42	industrial gases	63
43	science and technology in the soviet union	62
44	security technology	61
45	technological change	60
46	rail technologies	59
47	steam locomotive technologies	59
48	information technology	58
49	pollution control technologies	57
50	science and technology studies	55

Heuristics

Article Sample

Cable television headend	Iranian Science and Culture Hall of Fame	Pakistan Atomic Research Reactor
Do-it-yourself biology	GraphExeter	Cybathlon
Community technology	ARGUS-IS	Light-emitting diode
GESMES/TS	Timeline of music technology	Multivision (television technology)
Lidar	MMN80CPU	Timeline of Airbnb
IT chargeback and showback	Skype for Business Server	Waveform (podcast)
Resende Nuclear Fuel Factory	Consumption map	Video editing
Nintendo Gateway System	Natural Health Products Directorate	Timeline of Dropbox
Switched mesh	FLEPia	Active queue management
Religious response to assisted reproductive technology	Weissach axle	The Audacity to Podcast
Rubric (academic)	Dual-use technology	Axial piston pump
Ephemeralization	Phantom vibration syndrome	Microvesicles
Pasotron	Halo (safety device)	Nanoracks CubeSat Deployer
Capacity management	Overshoot (microwave communication)	Condenser (heat transfer)
Dynamic design analysis method	Freight bicycle	IBM System Object Model
Iranian missile tests	Graphical identification and authentication	TENET (network)
Powertec RPA	Spaceplane	Data center bridging
Magnesium/Teflon/Viton	Watering trough	Turbine
Logitech Harmony	Bash valve	Political Economy of Research and Innovation
Telesync	Water pyramid	Stroke ratio
Directed-energy weapon	Stream ripping	Bank clearing number
Air Force Network	Restrictive flow orifice	Radar
Winepress	Remote access policy	WebQuest
Advanced Train Management System	Hydrargyrum medium-arc iodide lamp	Third-generation sequencing
Thermotunnel cooling	Xenos (graphics chip)	Thermoelectric cooling
HARASSmap	Digital newspaper technology	Zen (portable media player)
EFx Factory	Ice Stupa	Quantum technology
MicroMasters	Microscope	Videophile
Transcranial direct-current stimulation	Space suit	Distributed File System (Microsoft)
Drilling stabilizer	Single-source data	Nanofountain probe

Outline

1. Heuristics

2. Classification

3. Field Study

Classification

Idea

- Articles with heading 'history' at top level as training data [Aprosio 2015]
  - heading 'history' → label HISTORY
  - any other heading → label OTHER
- 1,584 articles, 8,504 sections
  - 1,584 history sections (18.63%)
  - 6,920 other sections (81.37%)
- Feature selection:
  - unify/map years
  - unify/map persons
  - vocabulary size (document frequency)
  - binary or relative term frequency
  - oversampling
- 5-fold cross-validation on training data to find best classifier(s)
- Introduction excluded for training and finding best classifier(s)
- Test on articles without designated history sections (include intro in test)

Classification

Scikit-learn

- NEAREST NEIGHBOURS: K-Nearest Neighbors Classifier
- NAIVE BAYES: Bernoulli, Categorical, Complement, Gaussian, Multinomial
- DECISION TREES: Decision Tree Classifier, Decision Tree Regressor
- ENSEMBLE: AdaBoost, Bagging, Random Forest, Extra-Trees, Gradient Boosting, Histogram-based Gradient Boosting
- NEURAL NETWORKS: Multi-Level Perceptron
- LINEAR MODELS: Logistic Regression, Passive Aggressive Classifier, Perceptron, Ridge Classifier, Stochastic Gradient Descent, Stochastic Gradient Descent One-Class SVM
- QUADRATIC DISCRIMINANT: Quadratic Discriminant Analysis
- SUPPORT VECTOR MACHINES: Linear Kernel SVC, Poly Kernel SVC, Sigmoid Kernel SVC, Radial Base Function SVC

Classification

Results: Precision ≥ 0.75, Recall ≥ 0.45, sorted by Precision first, Recall second

	OVERSAMPLING	YEARS	PERSONS	VOCAB SIZE	TERMS	PRECISION	RECALL
RandomForestClassifier	0	1	0	1000	relative	0.866	0.481
ExtraTreesClassifier	0	0	0	1000	binary	0.860	0.459
ExtraTreesClassifier	0	1	0	1000	binary	0.858	0.476
RandomForestClassifier	0	1	1	1000	relative	0.856	0.498
ExtraTreesClassifier	0	0	1	1000	binary	0.855	0.457
ExtraTreesClassifier	0	1	1	1000	binary	0.844	0.456
ExtraTreesClassifier	0	1	1	100	relative	0.832	0.493
RBFSupportVectorClassifier	0	0	1	1000	binary	0.832	0.482
ExtraTreesClassifier	0	1	0	100	relative	0.830	0.499
RBFSupportVectorClassifier	0	1	1	1000	binary	0.829	0.487
...	...	...	...	...	...	...	...
GradientBoostingClassifier	0	0	1	1000	binary	0.809	0.538
...	...	...	...	...	...	...	...
MultiLevelPerceptronClassifier	0	0	0	10000	binary	0.763	0.613

Classification

Random Forest Classifier

- 2,825 articles w/o heading with 'histor[y|i]': ~375 with history section (13.24%)
- 1,584 articles with heading 'history': ~1,569 with history section (99.03%)
- Heuristic only:
  - Precision ≈ 0.990
  - Recall = 1,569 / (375 + 1,569) ≈ 0.807
- Heuristic first, Random Forest Classifier (P = 0.866, R = 0.481) as fallback:
  - Precision = 1,749 / (43 + 1,749) ≈ 0.976
  - Recall = 1,749 / (375 + 1,569) ≈ 0.900

  (TP = 1,749 (1,569 from H + 180 from O), FP = 43 (15 from H + 28 from O))

Classification

Extra-Trees Classifier

- 2,825 articles w/o heading with 'histor[y|i]': ~375 with history section (13.24%)
- 1,584 articles with heading 'history': ~1,569 with history section (99.03%)
- Heuristic only:
  - Precision ≈ 0.990
  - Recall = 1,569 / (375 + 1,569) ≈ 0.807
- Heuristic first, Extra-Trees Classifier (P = 0.860, R = 0.459) as fallback:
  - Precision = 1,741 / (43 + 1,741) ≈ 0.976
  - Recall = 1,741 / (375 + 1,569) ≈ 0.896

  (TP = 1,741 (1,569 from H + 172 from O), FP = 43 (15 from H + 28 from O))

Classification

RBF Support Vector Machine

- 2,825 articles w/o heading with 'histor[y|i]': ~375 with history section (13.24%)
- 1,584 articles with heading 'history': ~1,569 with history section (99.03%)
- Heuristic only:
  - Precision ≈ 0.990
  - Recall = 1,569 / (375 + 1,569) ≈ 0.807
- Heuristic first, RBF Support Vector Machine (P = 0.832, R = 0.482) as fallback:
  - Precision = 1,749 / (51 + 1,749) ≈ 0.972
  - Recall = 1,749 / (375 + 1,569) ≈ 0.900

  (TP = 1,749 (1,569 from H + 180 from O), FP = 51 (15 from H + 36 from O))

Classification

GradientBoostingClassifier

- 2,825 articles w/o heading with 'histor[y|i]': ~375 with history section (13.24%)
- 1,584 articles with heading 'history': ~1,569 with history section (99.03%)
- Heuristic only:
  - Precision ≈ 0.990
  - Recall = 1,569 / (375 + 1,569) ≈ 0.807
- Heuristic first, Gradient Boosting Classifier (P = 0.809, R= 0.538) as fallback:
  - Precision = 1,770 / (62 + 1,770) ≈ 0.966
  - Recall = 1,770 / (375 + 1,569) ≈ 0.910

  (TP = 1,770 (1,569 from H + 201 from O), FP = 62 (15 from H + 47 from O))

Classification

MultiLevelPerceptronClassifier

- 2,825 articles w/o heading with 'histor[y|i]': ~375 with history section (13.24%)
- 1,584 articles with heading 'history': ~1,569 with history section (99.03%)
- Heuristic only:
  - Precision ≈ 0.990
  - Recall = 1,569 / (375 + 1,569) ≈ 0.807
- Heuristic first, Multi-Level Perceptron Classifier (Precision = 0.763, Recall = 0.613) as fallback:
  - Precision = 1,798 / (86 + 1,798) ≈ 0.954
  - Recall = 1,798 / (375 + 1,569) ≈ 0.923

    (TP = 1,798 (1,569 from H + 229 from O), FP = 86 (15 from H + 71 from O))

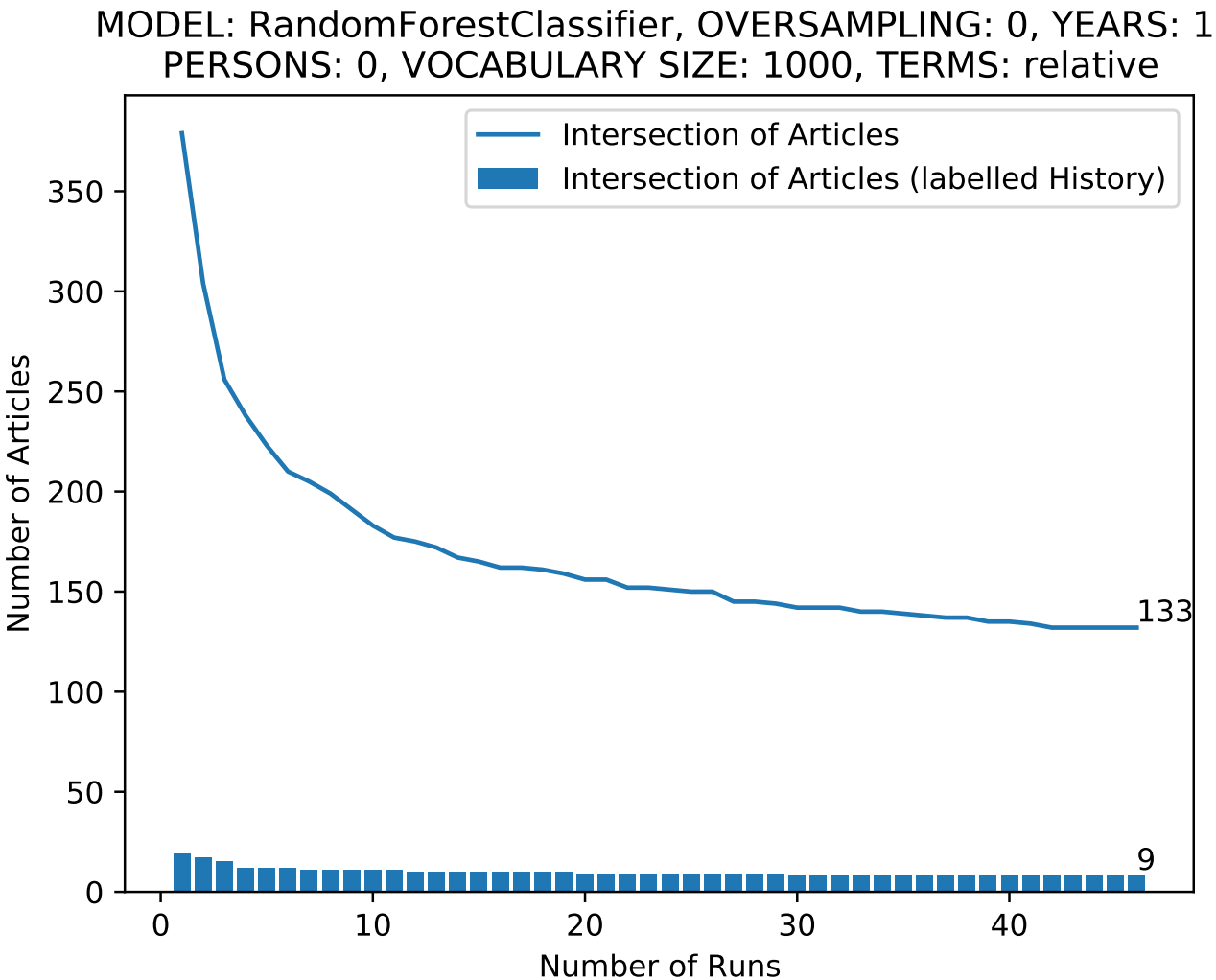
Classification

Trial & Issues

- Trial above classifiers against articles without designated history sections
- Some classifiers non-deterministic: repeat, get intersection of articles of all runs, repeat until size of intersection does not decrease in five runs

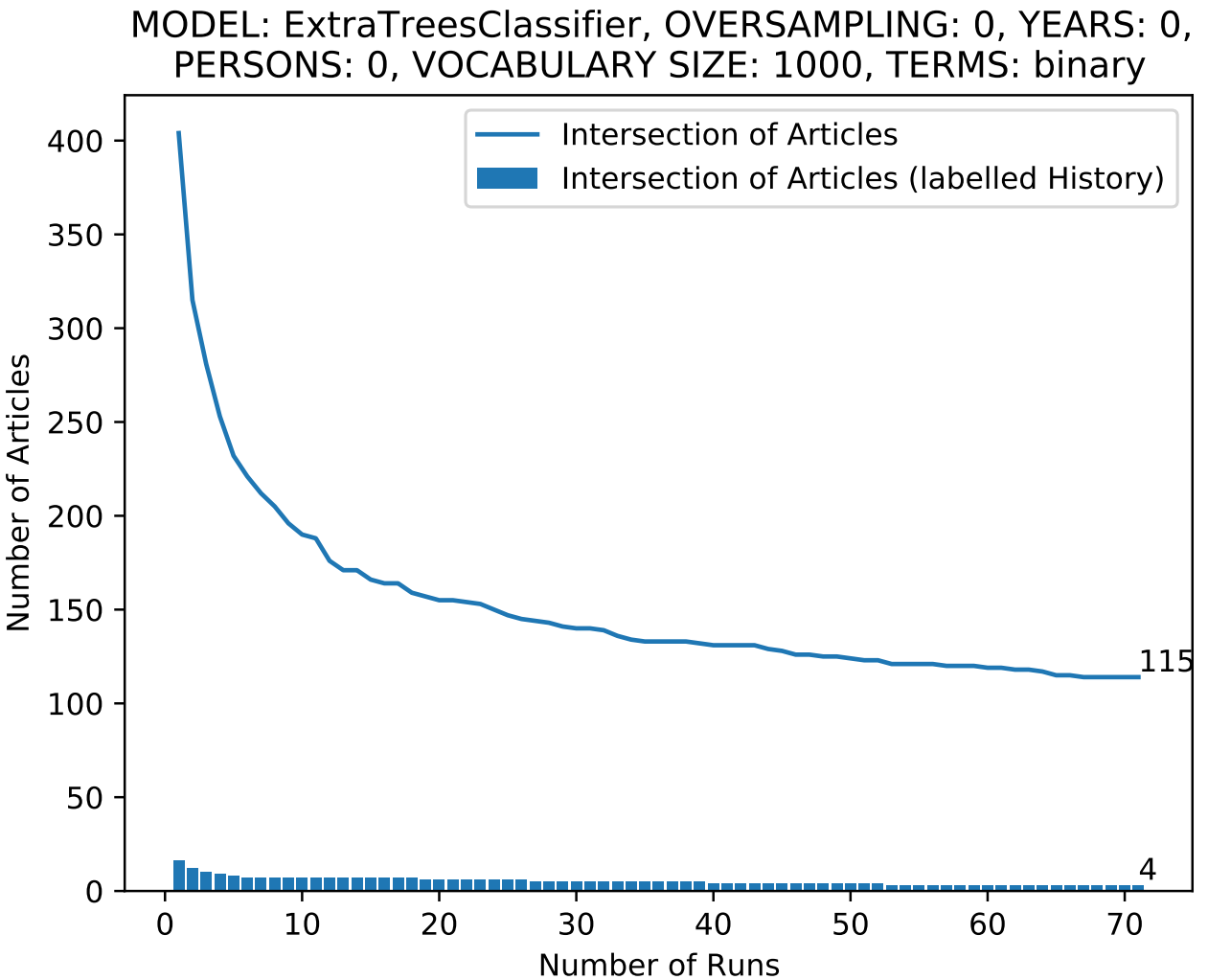
Classification

Random Forest Classifier



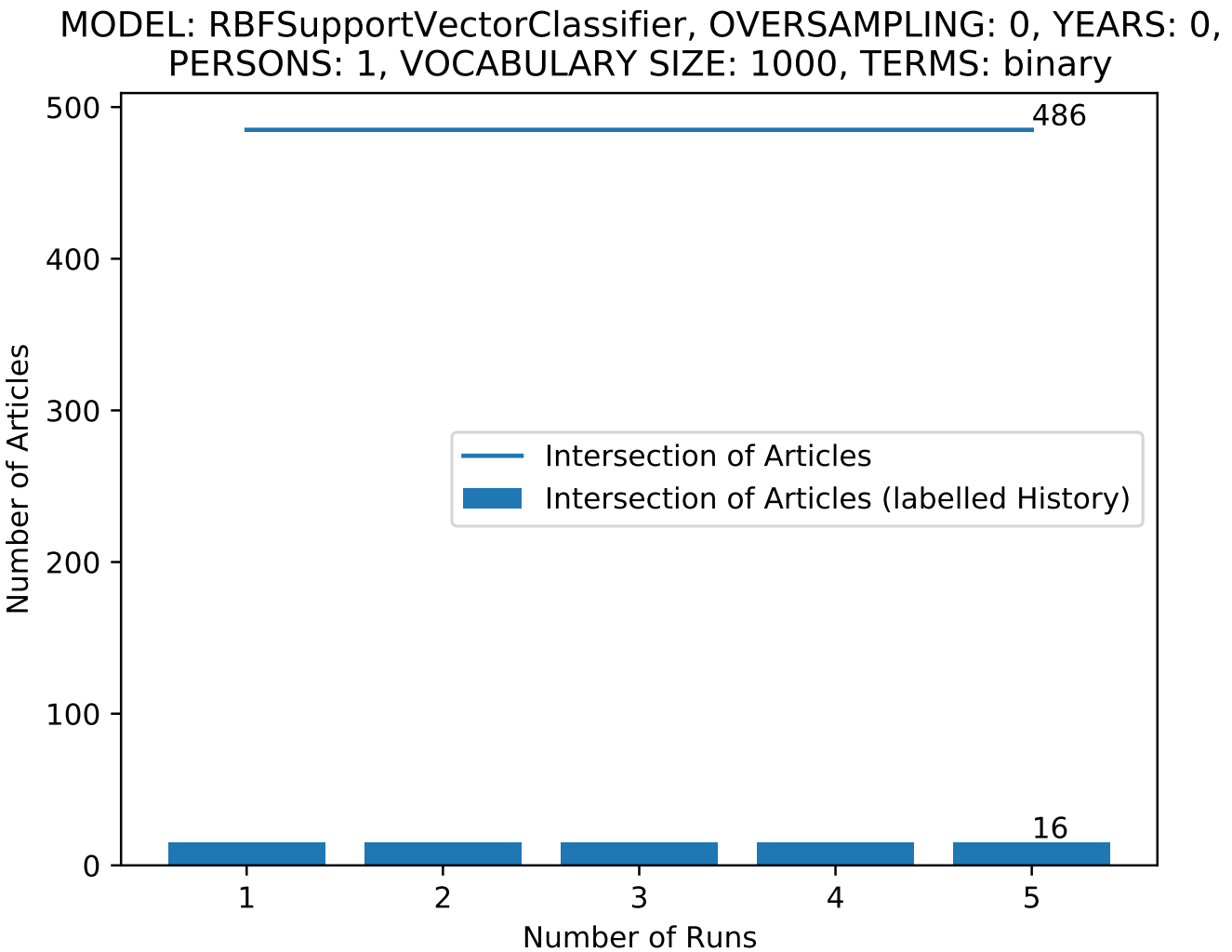
Classification

Extra-Trees Classifier



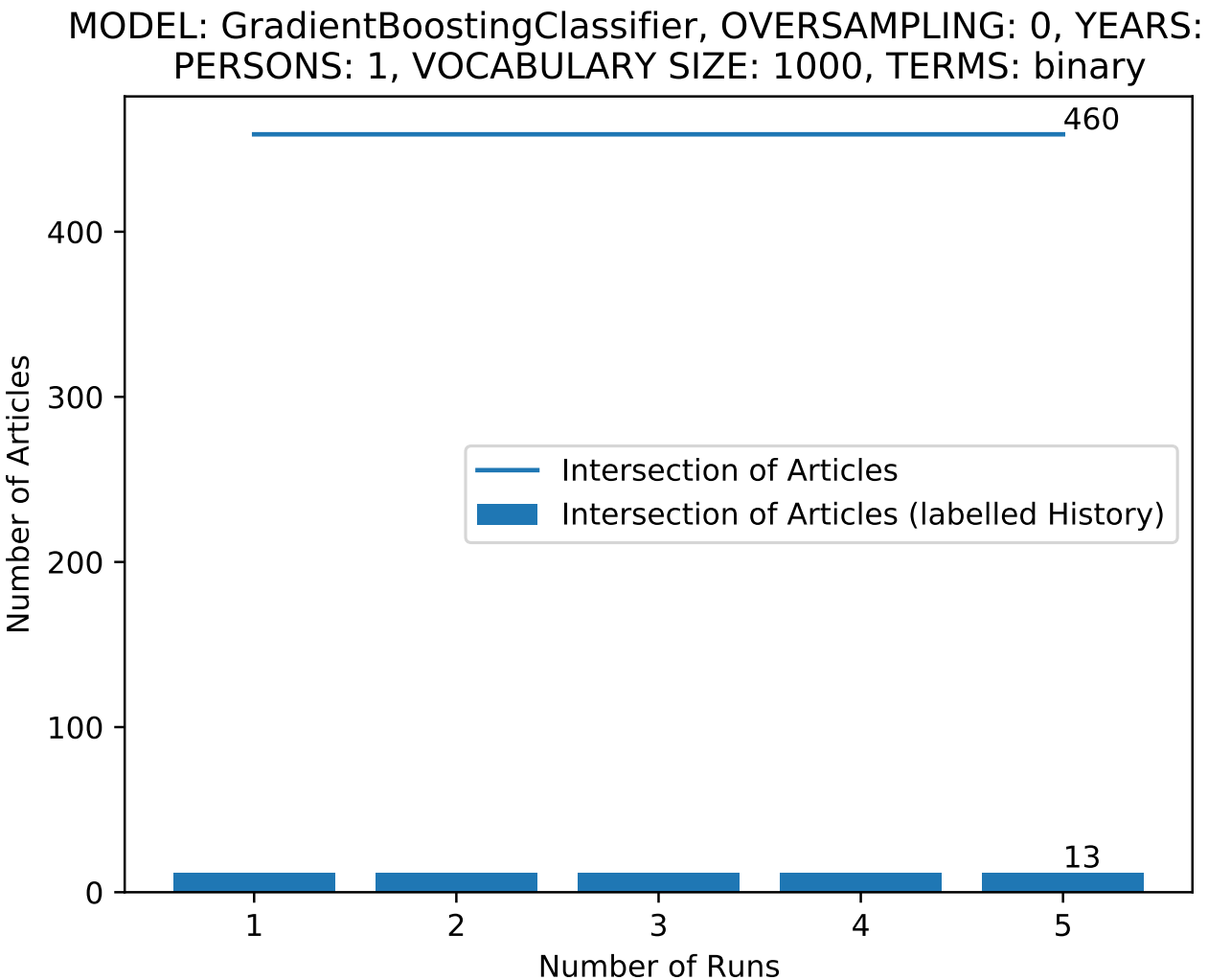
Classification

RBF Support Vector Machine



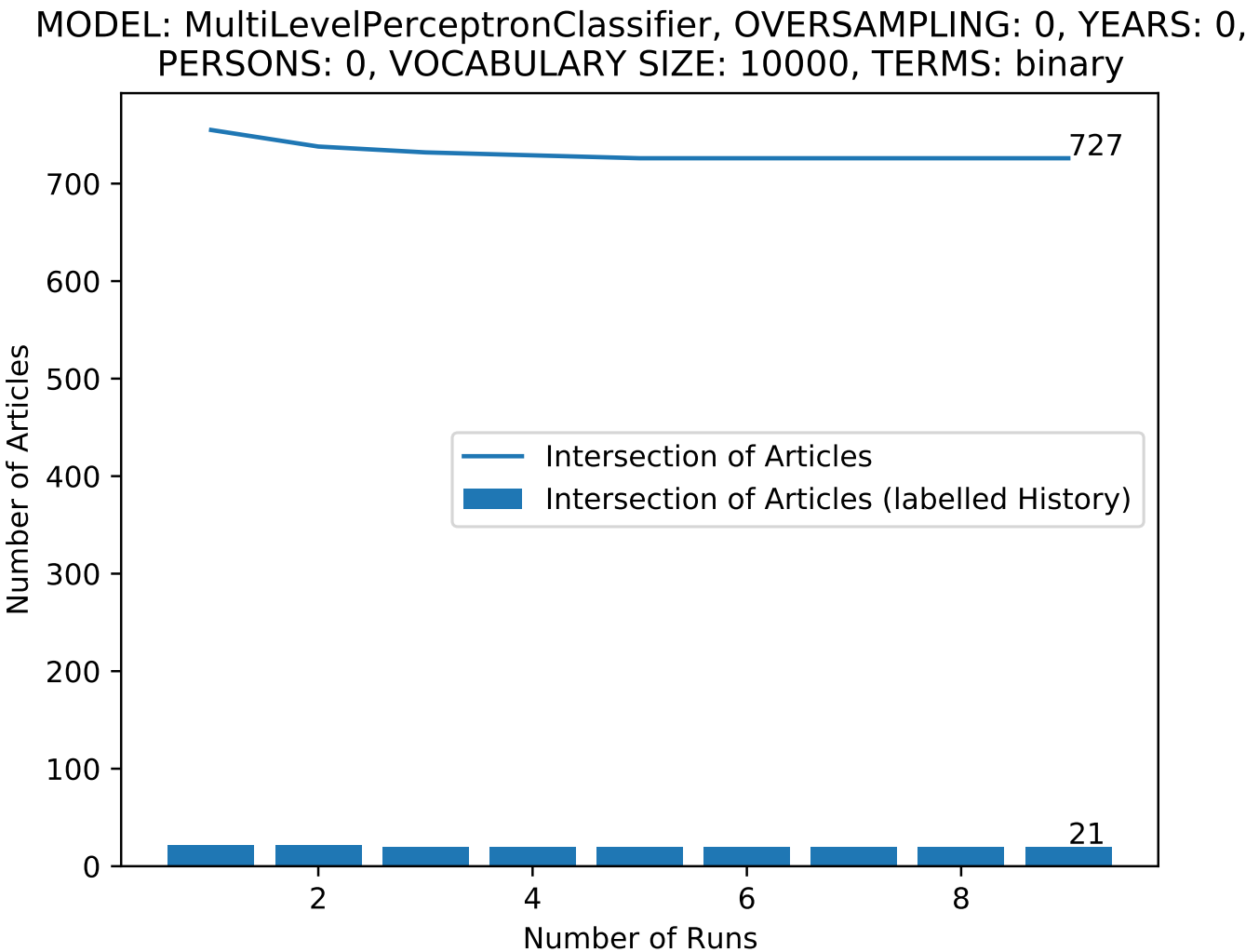
Classification

GradientBoostingClassifier



Classification

MultiLevelPerceptronClassifier



Classification

Trial & Issues

- Trial above classifiers against articles without designated history sections
- Some classifiers non-deterministic: repeat, get intersection of articles of all runs, repeat until size of intersection does not decrease in five runs
- Issue: first setup did not include sections with less than 100 characters
- Model selection excluding sections with less than 100 charcters with slightly worse precision and/or recall for some classifiers

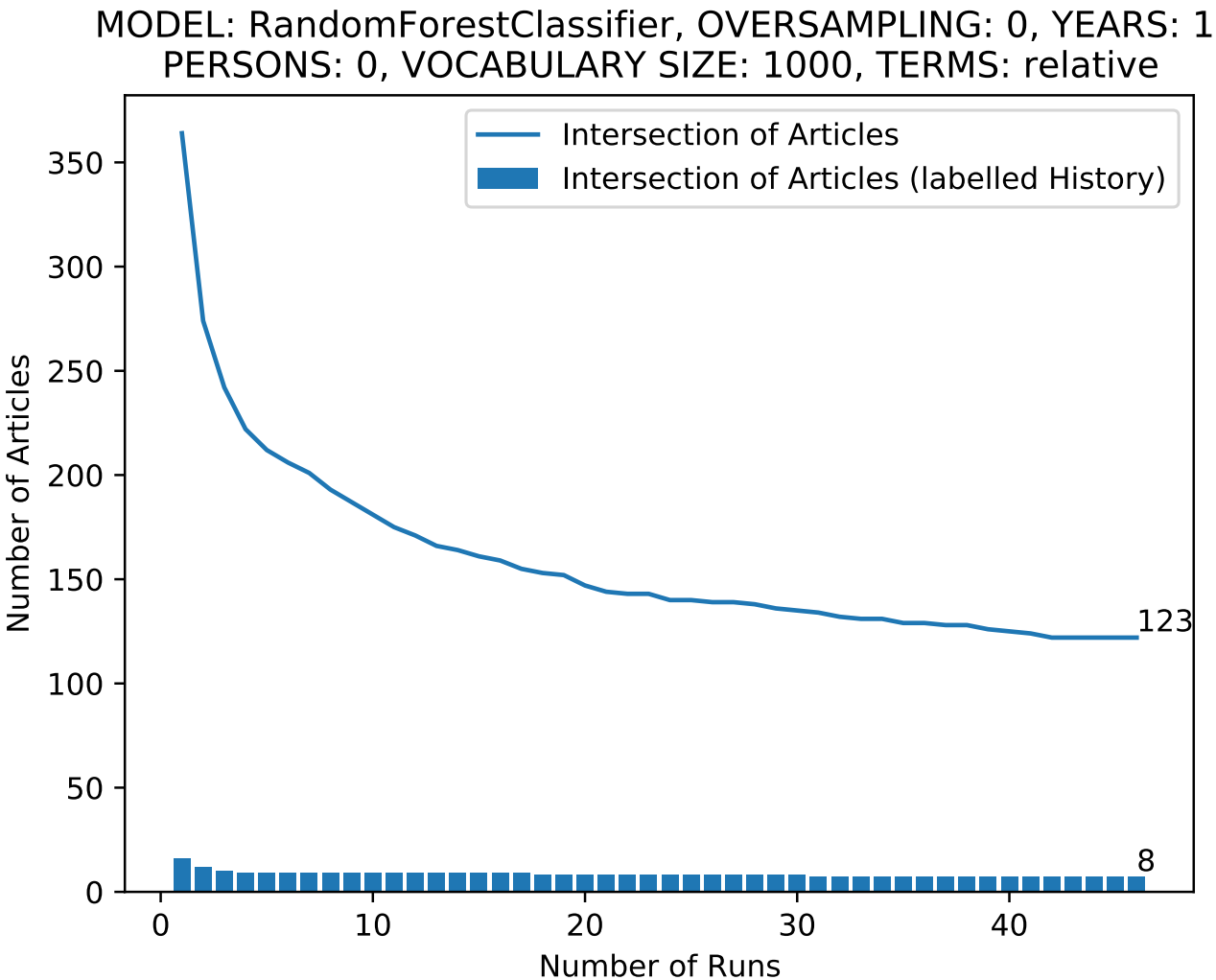
Classification

Results: Precision ≥ 0.75, Recall ≥ 0.45, sorted by Precision first, Recall second

	OVERSAMPLING	YEARS	PERSONS	VOCAB SIZE	TERMS	PRECISION	RECALL
ExtraTreesClassifier	0	0	0	1000	binary	0.861	0.450
RandomForestClassifier	0	1	0	1000	relative	0.857	0.482
ExtraTreesClassifier	0	1	1	1000	binary	0.855	0.461
RandomForestClassifier	0	1	1	1000	relative	0.848	0.483
ExtraTreesClassifier	0	1	0	1000	binary	0.846	0.464
ExtraTreesClassifier	0	1	1	100	relative	0.845	0.495
RBFSupportVectorClassifier	0	0	1	1000	binary	0.832	0.500
RBFSupportVectorClassifier	0	1	1	1000	binary	0.831	0.498
...	...	...	...	...	...	...	...
GradientBoostingClassifier	0	1	0	1000	binary	0.805	0.546
GradientBoostingClassifier	0	0	0	10000	relative	0.805	0.543
...	...	...	...	...	...	...	...
MultiLevelPerceptronClassifier	0	1	1	10000	binary	0.761	0.598
...	...	...	...	...	...	...	...

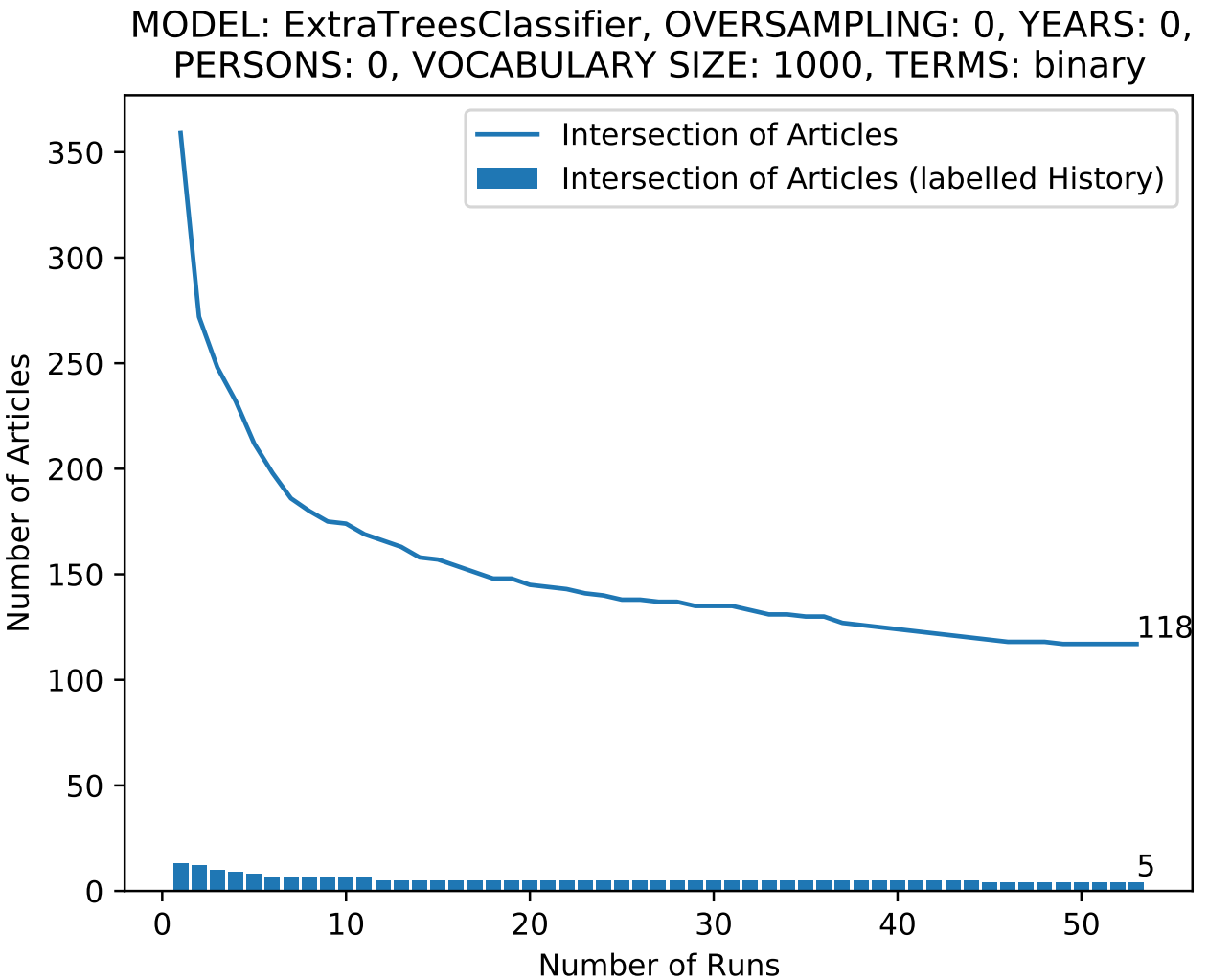
Classification

Random Forest Classifier



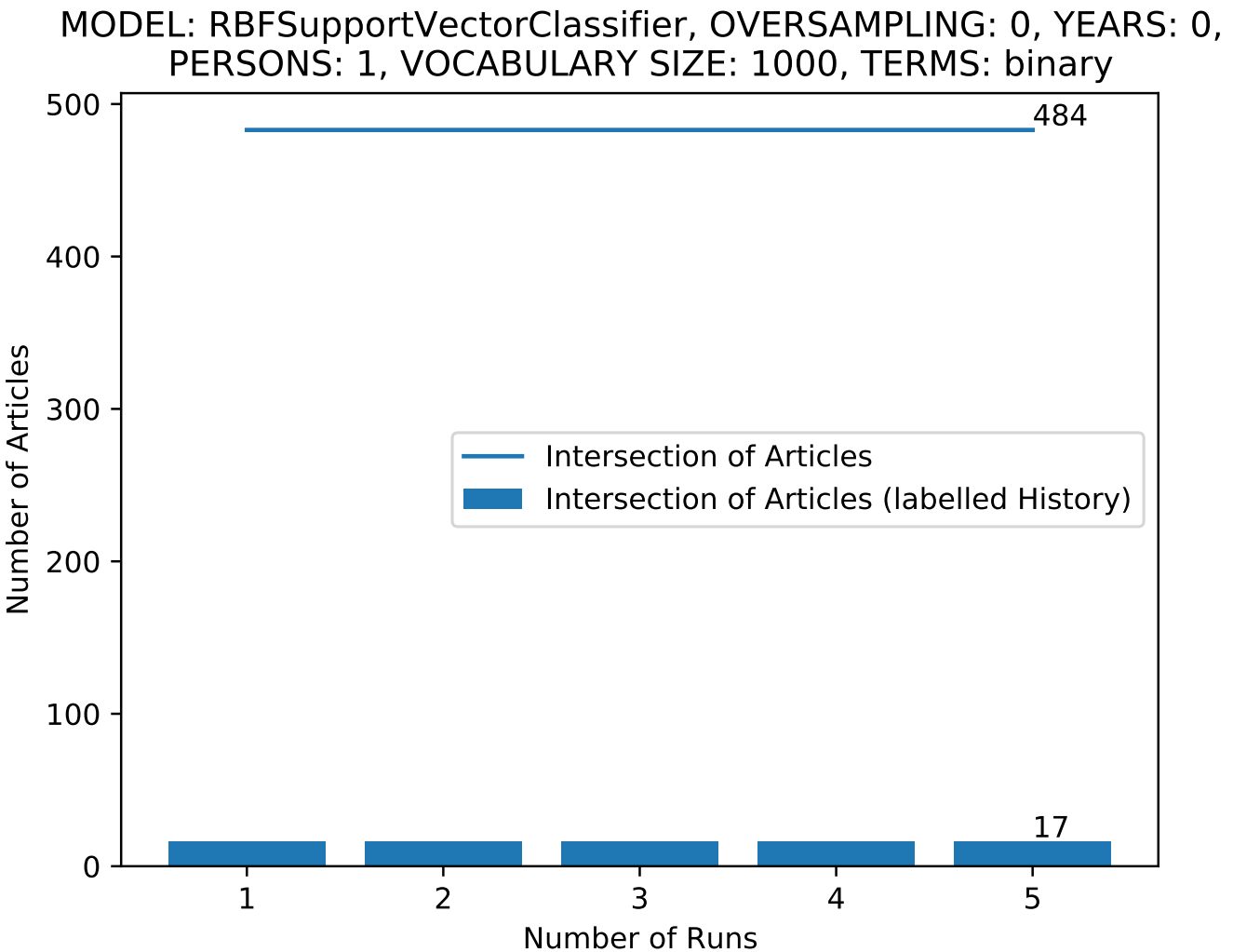
Classification

Extra-Trees Classifier



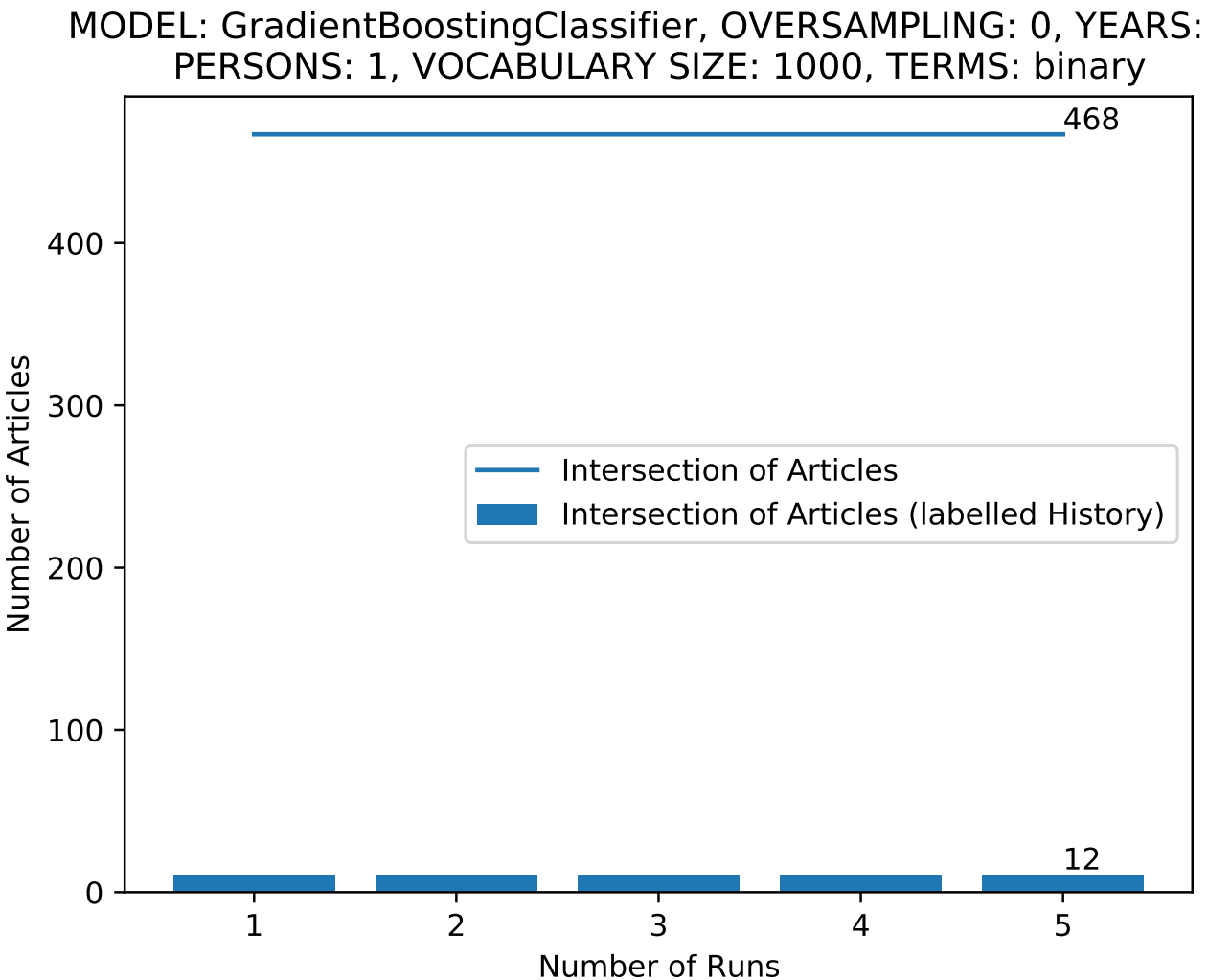
Classification

RBF Support Vector Machine



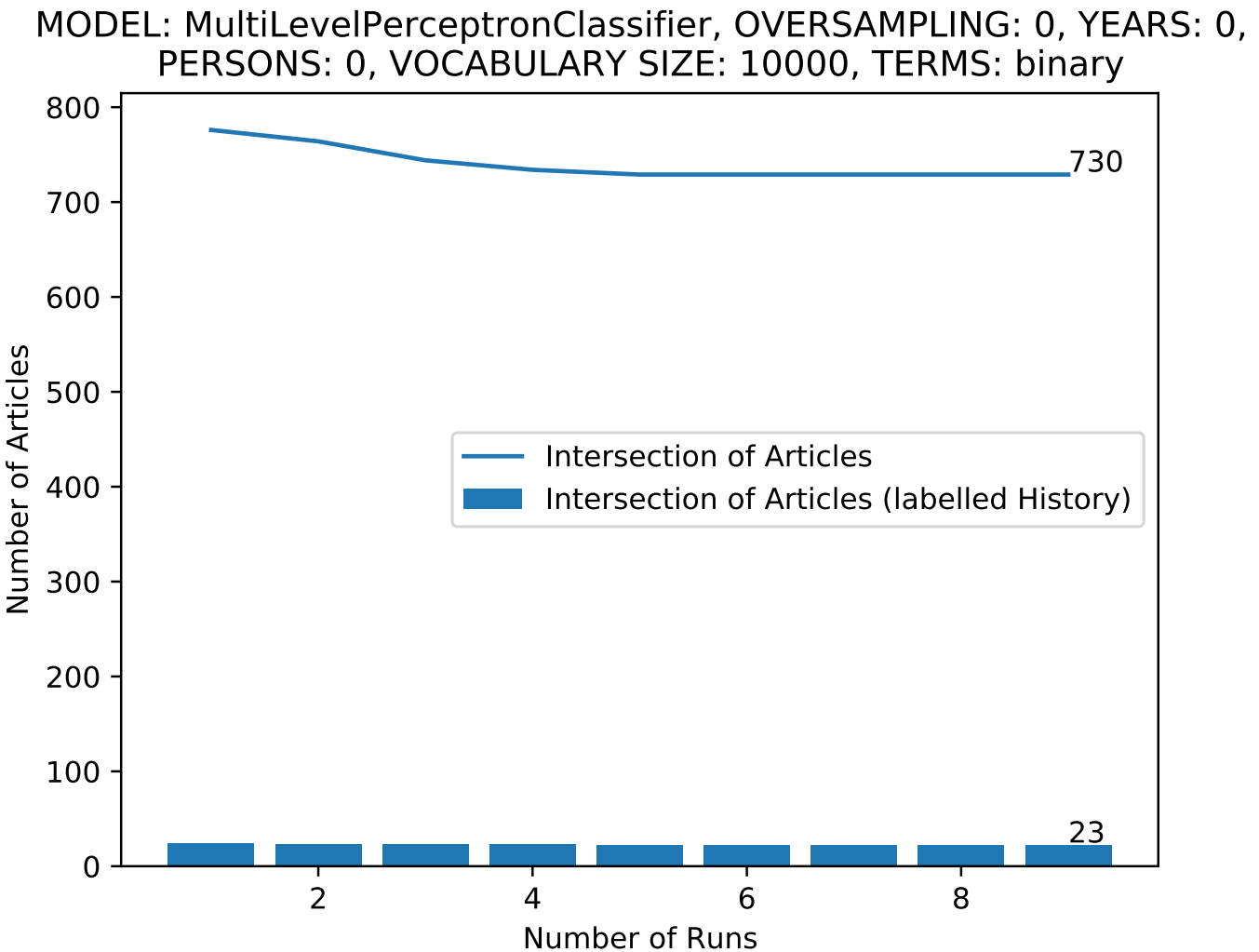
Classification

GradientBoostingClassifier



Classification

MultiLevelPerceptronClassifier



Classification

Trial & Issues

- Trial above classifiers against articles without designated history sections
- Some classifiers non-deterministic: repeat, get intersection of articles of all runs, repeat until size of intersection does not decrease in five runs
- Issue: first setup did not include sections with less than 100 characters
- Model selection excluding sections with less than 100 charcters with slightly worse precision and/or recall for some classifiers
- Model selection excluding sections with less than 100 charcters yielded one more article from those determined to contain history section during first evaluation (29 instead of 28)
- BERT:
  - including sections with $<$ 100 characters: 393 articles, incl. 14 from those determined to contain history section during first evaluation, Precision = 0.728 and Recall = 0.347
  - excluding sections with $<$ 100 characters: 369 articles, incl. 15 from those determined to contain history section during first evaluation, Precision = 0.811 and Recall = 0.368

Classification

Evaluation

- 1013 articles in evaluation pool
- 8 batches with 100 articles, 1 batch (010) with 128 (contains 29 articles from first evaluation) from Sklearn trials
- 8 labellers (7 out of 9 of first run, 1 new labeller for batch 001, no batch 007)
  - Labeller 001: 55 labelled as having a history section (54%), 5 without label
  - Labeller 002: 62 labelled as having a history section (62%), 0 without label
  - Labeller 003: 82 labelled as having a history section (82%), 2 without label
  - Labeller 004: 0 labelled as having a history section (0%), 27 without label
    $\rightarrow$ labelled again: 56 labelled as having a history section (56%), 1 without label
  - Labeller 005: 76 labelled as having a history section (82%), 4 without label
  - Labeller 006: 61 labelled as having a history section (61%), 6 without label
  - Labeller 008: 37 labelled as having a history section (37%), 0 without label
  - Labeller 009: 55 labelled as having a history section (55%), 0 without label
  - Labeller 010: 64 labelled as having a history section (65%), 0 without label
- No inter-labeller agreement; batch 005 (partially), 009, 010 labelled by me
- Batches 011 (68) and 012 (17) for additional BERT articles labelled by me

Classification

Evaluation

- Precision and recall (615 articles labelled to contain history section):

	T		1		2		3		4		5	
	P	R	P	R	P	R	P	R	P	R	P	R
RandomForestClassifier	0.86	0.48	0.82	0.16	0.72	0.14	0.54	0.11	0.30	0.06	0.68	0.14
ExtraTreesClassifier	0.86	0.45	0.73	0.14	0.64	0.12	0.40	0.08	0.25	0.05	0.55	0.12
RBFSupportVectorClassifier	0.83	0.50	0.61	0.48	0.59	0.46	0.52	0.41	0.29	0.23	0.57	0.36
GradientBoostingClassifier	0.81	0.55	0.60	0.45	0.57	0.43	0.48	0.37	0.29	0.22	0.54	0.35
MultiLevelPerceptronClassifier	0.76	0.60	0.51	0.61	0.46	0.55	0.35	0.42	0.20	0.23	0.43	0.48
BERT	0.81	0.37	0.59	0.35	0.55	0.33	0.46	0.28	0.22	0.13	0.52	0.26

T - as per test run on articles with designated history section

1 - at least one section labelled by classifier labelled by labeller:

classifier: A (46x), B (10x) vs labeller: B, C

2 - section name most frequently labelled by classifier labelled by labeller:

classifier: A (46x), B (10x) vs labeller: A, C

3 - classifier section names is subset of labeller section names:

classifier: A (46x), B (10x) vs labeller: A, B, C

4 - classifier and labeller labelled same sections:

classifier: A (46x), B (10x) vs labeller: A, B

5 - mean (sum of precision/recall of each article devided by number of articles/615):

classifier: A (46x), B (10x), E (7x), F (5x) vs labeller: A, B, C, D $\rightarrow$ P = 0.5, R = 0.5

Conclusion

- Heuristics sufficient to find science & technology articles
- Classifier required to find more history sections
- Evaluation yields mixed results
- Choice of P/R estimation task-dependent (high recall for manual check, high precision for automation)