UniNE at CLEF 2016: Author Clustering

Mirco Kocher

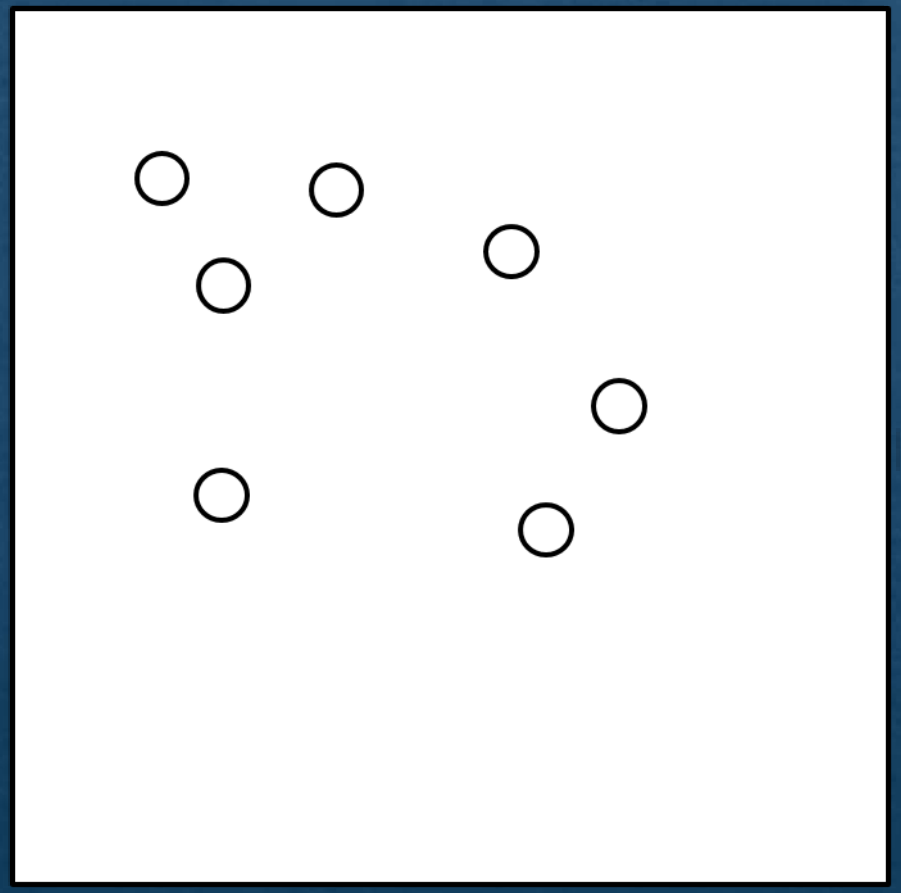
University of Neuchâtel, Switzerland

Task – Description

- Collection of up to 100 documents
- Identify authorship links and groups of documents by the same author
- All documents are single-authored
- Same language (Dutch, English, or Greek)
- Same genre (newspaper articles or reviews)
- Topic or text length may vary
- Number of authors in collection not known

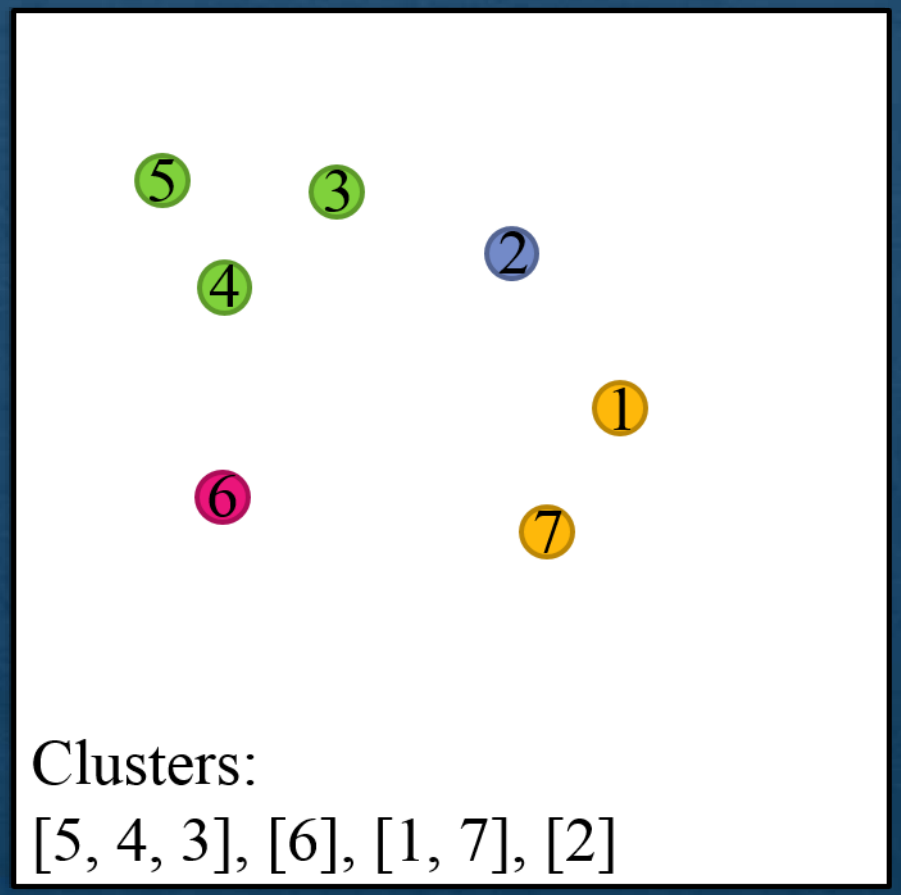
UNIVERSITÉ DE NEUCHÂTEL

Task – Input



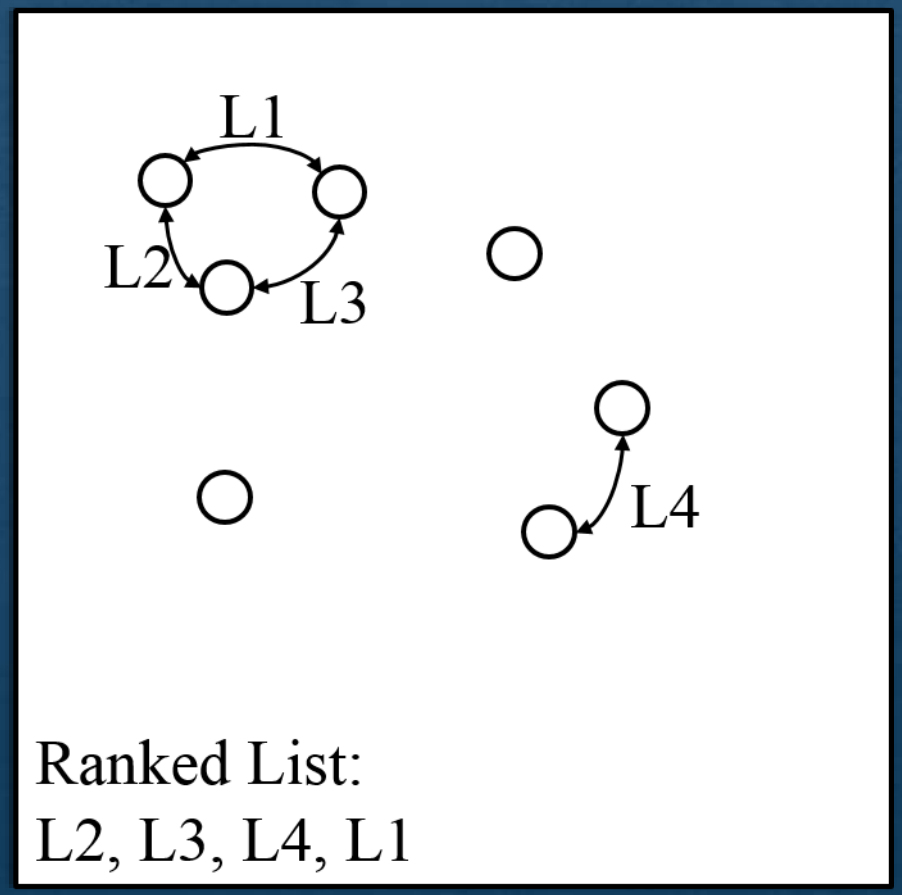
UNIVERSITÉ DE NEUCHÂTEL

Task – Output



Clusters:
[5, 4, 3], [6], [1, 7], [2]

+



Ranked List:
L2, L3, L4, L1

Training Data

- Text length variation
  - Dutch reviews: only 130 token per review
  - Dutch newspaper: over 1.100 token per article
- Overall: many different authors
  - Greek reviews contains 55 texts written by 50 authors
  - English reviews, 62 out of 70 authors have written only a single document

Evaluation

- Clustering: BCubed F-Score
  - Harmonic mean of precision and recall for each document
  - Document precision: noise in cluster?
  - Document recall: complete cluster?
- Ranking: Mean Average Precision
  - Approx. area under precision-recall curve
  - Clear emphasis on first position
  - Misclassification with low probability is less penalized

UNIVERSITÉ DE NEUCHÂTEL

Our Baseline

- One text = one cluster
  - Document precision = 100%
  - Document recall is lower, but not many big clusters expected
- Random scores for all combinations
  - MAP can only increase

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Features Selection

- Extract top *m* most frequent terms
  - Isolated words (no stemming)
  - Punctuation symbols
- *m* at most 200, was usually below
  - Rather short documents
  - Without words appearing only once

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Distance

$$\Delta(A, B) = \sum_{i=1}^{m} |P_A[f_i] - P_B[f_i]|$$

- Manhattan distance from document A to document B
- Vectors with relative frequencies of features
- Vector from document B according to m features of document A
  - Not symmetric

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Distance Matrix

	1	4	6	9	17	22	23	24	27	28	36	42	43	47
1		303	353	312	269	245	307	353	298	262	340	297	307	297
4	299		340	338	310	286	314	344	377	300	333	321	303	307
6	330	299		325	305	306	294	310	352	271	304	278	322	268
9	246	279	290		273	265	290	290	290	251	265	280	231	252
17	284	309	363	335		284	353	373	347	318	360	335	315	327
22	264	307	354	340	268		303	330	326	265	324	315	321	271
23	321	337	338	310	344	303		331	379	295	348	338	304	291
24	408	412	429	359	442	384	349		442	364	379	409	358	331
27	307	365	362	348	320	340	344	354		328	371	336	333	350
28	279	300	323	321	289	247	271	351	355		345	307	323	275
36	326	326	342	318	312	290	311	297	351	314		332	278	271
42	239	267	277	300	262	230	250	279	266	256	279		288	249
43	341	344	391	285	337	342	313	312	378	341	334	382		311
47	308	299	320	322	317	271	275	298	360	276	304	327	284	

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Indication H

- Mean distance of A to all other documents, mean(A, X)
- Standard deviation of A to all other documents, SD(A, X)
- Check if: $\Delta(A, B) \leq \text{mean}(A, X) - 2.0 * \text{SD}(A, X)$
  - Horizontal indication for authorship link

Our Approach – Indication H

	1	4	6	9	17	22	23	24	27	28	36	42	43	47		μ	σ	μ-2.0*σ
1		303	353	312	269	245	307	353	298	262	340	297	307	297		303	31	241
4	299		340	338	310	286	314	344	377	300	333	321	303	307		321	24	274
6	330	299		325	305	306	294	310	352	271	304	278	322	268		305	23	259
9	246	279	290		273	265	290	290	290	251	265	280	231	252		269	19	232
17	284	309	363	335		284	353	373	347	318	360	335	315	327		331	27	276
22	264	307	354	340	268		303	330	326	265	324	315	321	271		307	29	248
23	321	337	338	310	344	303		331	379	295	348	338	304	291		326	24	278
24	408	412	429	359	442	384	349		442	364	379	409	358	331		390	35	319
27	307	365	362	348	320	340	344	354		328	371	336	333	350		343	18	308
28	279	300	323	321	289	247	271	351	355		345	307	323	275		307	32	242
36	326	326	342	318	312	290	311	297	351	314		332	278	271		313	23	267
42	239	267	277	300	262	230	250	279	266	256	279		288	249		265	19	226
43	341	344	391	285	337	342	313	312	378	341	334	382		311		339	29	280
47	308	299	320	322	317	271	275	298	360	276	304	327	284			305	24	256

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Indication H

	1	4	6	9	17	22	23	24	27	28	36	42	43	47	μ	σ	μ-2.0*σ
1		303	353	312	269	245	307	353	298	262	340	297	307	297	303	31	241
4	299		340	338	310	286	314	344	377	300	333	321	303	307	321	24	274
6	330	299		325	305	306	294	310	352	271	304	278	322	268	305	23	259
9	246	279	290		273	265	290	290	290	251	265	280	231	252	269	19	232
17	284	309	363	335		284	353	373	347	318	360	335	315	327	331	27	276
22	264	307	354	340	268		303	330	326	265	324	315	321	271	307	29	248
23	321	337	338	310	344	303		331	379	295	348	338	304	291	326	24	278
24	408	412	429	359	442	384	349		442	364	379	409	358	331	390	35	319
27	307	365	362	348	320	340	344	354		328	371	336	333	350	343	18	308
28	279	300	323	321	289	247	271	351	355		345	307	323	275	307	32	242
36	326	326	342	318	312	290	311	297	351	314		332	278	271	313	23	267
42	239	267	277	300	262	230	250	279	266	256	279		288	249	265	19	226
43	341	344	391	285	337	342	313	312	378	341	334	382		311	339	29	280
47	308	299	320	322	317	271	275	298	360	276	304	327	284		305	24	256

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Indication V

- Mean distance of all other documents to B, mean(X, B)
- Standard deviation of all other documents to B, SD(X, B)
- Check if: $\Delta(A, B) \leq \text{mean}(X, B) - 2.0 * \text{SD}(X, B)$
  - Vertical indication for authorship link

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Indication V

	1	4	6	9	17	22	23	24	27	28	36	42	43	47
1		303	353	312	269	245	307	353	298	262	340	297	307	297
4	299		340	338	310	286	314	344	377	300	333	321	303	307
6	330	299		325	305	306	294	310	352	271	304	278	322	268
9	246	279	290		273	265	290	290	290	251	265	280	231	252
17	284	309	363	335		284	353	373	347	318	360	335	315	327
22	264	307	354	340	268		303	330	326	265	324	315	321	271
23	321	337	338	310	344	303		331	379	295	348	338	304	291
24	408	412	429	359	442	384	349		442	364	379	409	358	331
27	307	365	362	348	320	340	344	354		328	371	336	333	350
28	279	300	323	321	289	247	271	351	355		345	307	323	275
36	326	326	342	318	312	290	311	297	351	314		332	278	271
42	239	267	277	300	262	230	250	279	266	256	279		288	249
43	341	344	391	285	337	342	313	312	378	341	334	382		311
47	308	299	320	322	317	271	275	298	360	276	304	327	284	

μ	304	319	345	324	311	292	306	325	348	295	330	327	305	292
σ	43	37	38	19	46	42	30	28	44	34	33	35	30	30
μ-2.0*σ	218	245	269	285	220	208	247	269	260	227	264	257	246	232

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Indication V

	1	4	6	9	17	22	23	24	27	28	36	42	43	47
1		303	353	312	269	245	307	353	298	262	340	297	307	297
4	299		340	338	310	286	314	344	377	300	333	321	303	307
6	330	299		325	305	306	294	310	352	271	304	278	322	268
9	246	279	290		273	265	290	290	290	251	265	280	231	252
17	284	309	363	335		284	353	373	347	318	360	335	315	327
22	264	307	354	340	268		303	330	326	265	324	315	321	271
23	321	337	338	310	344	303		331	379	295	348	338	304	291
24	408	412	429	359	442	384	349		442	364	379	409	358	331
27	307	365	362	348	320	340	344	354		328	371	336	333	350
28	279	300	323	321	289	247	271	351	355		345	307	323	275
36	326	326	342	318	312	290	311	297	351	314		332	278	271
42	239	267	277	300	262	230	250	279	266	256	279		288	249
43	341	344	391	285	337	342	313	312	378	341	334	382		311
47	308	299	320	322	317	271	275	298	360	276	304	327	284	

μ	304	319	345	324	311	292	306	325	348	295	330	327	305	292
σ	43	37	38	19	46	42	30	28	44	34	33	35	30	30
μ-2.0*σ	218	245	269	285	220	208	247	269	260	227	264	257	246	232

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Indication H & V

	1	4	6	9	17	22	23	24	27	28	36	42	43	47
1		303	353	312	269	245	307	353	298	262	340	297	307	297
4	299		340	338	310	286	314	344	377	300	333	321	303	307
6	330	299		325	305	306	294	310	352	271	304	278	322	268
9	246	279	290		273	265	290	290	290	251	265	280	231	252
17	284	309	363	335		284	353	373	347	318	360	335	315	327
22	264	307	354	340	268		303	330	326	265	324	315	321	271
23	321	337	338	310	344	303		331	379	295	348	338	304	291
24	408	412	429	359	442	384	349		442	364	379	409	358	331
27	307	365	362	348	320	340	344	354		328	371	336	333	350
28	279	300	323	321	289	247	271	351	355		345	307	323	275
36	326	326	342	318	312	290	311	297	351	314		332	278	271
42	239	267	277	300	262	230	250	279	266	256	279		288	249
43	341	344	391	285	337	342	313	312	378	341	334	382		311
47	308	299	320	322	317	271	275	298	360	276	304	327	284	

	1	4	6	9	17	22	23	24	27	28	36	42	43	47
1		303	353	312	269	245	307	353	298	262	340	297	307	297
4	299		340	338	310	286	314	344	377	300	333	321	303	307
6	330	299		325	305	306	294	310	352	271	304	278	322	268
9	246	279	290		273	265	290	290	290	251	265	280	231	252
17	284	309	363	335		284	353	373	347	318	360	335	315	327
22	264	307	354	340	268		303	330	326	265	324	315	321	271
23	321	337	338	310	344	303		331	379	295	348	338	304	291
24	408	412	429	359	442	384	349		442	364	379	409	358	331
27	307	365	362	348	320	340	344	354		328	371	336	333	350
28	279	300	323	321	289	247	271	351	355		345	307	323	275
36	326	326	342	318	312	290	311	297	351	314		332	278	271
42	239	267	277	300	262	230	250	279	266	256	279		288	249
43	341	344	391	285	337	342	313	312	378	341	334	382		311
47	308	299	320	322	317	271	275	298	360	276	304	327	284	

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Indication H &V

	1	4	6	9	17	22	23	24	27	28	36	42	43	47
1	0	303	353	312	269	245	307	353	298	262	340	297	307	297
4	299	0	340	338	310	286	314	344	377	300	333	321	303	307
6	330	299	0	325	305	306	294	310	352	271	304	278	322	268
9	246	279	290	0	273	265	290	290	290	251	265	280	231	252
17	284	309	363	335	0	284	353	373	347	318	360	335	315	327
22	264	307	354	340	268	0	303	330	326	265	324	315	321	271
23	321	337	338	310	344	303	0	331	379	295	348	338	304	291
24	408	412	429	359	442	384	349	0	442	364	379	409	358	331
27	307	365	362	348	320	340	344	354	0	328	371	336	333	350
28	279	300	323	321	289	247	271	351	355	0	345	307	323	275
36	326	326	342	318	312	290	311	297	351	314	0	332	278	271
42	239	267	277	300	262	230	250	279	266	256	279	0	288	249
43	341	344	391	285	337	342	313	312	378	341	334	382	0	311
47	308	299	320	322	317	271	275	298	360	276	304	327	284	0

UNIVERSITÉ DE NEUCHÂTEL

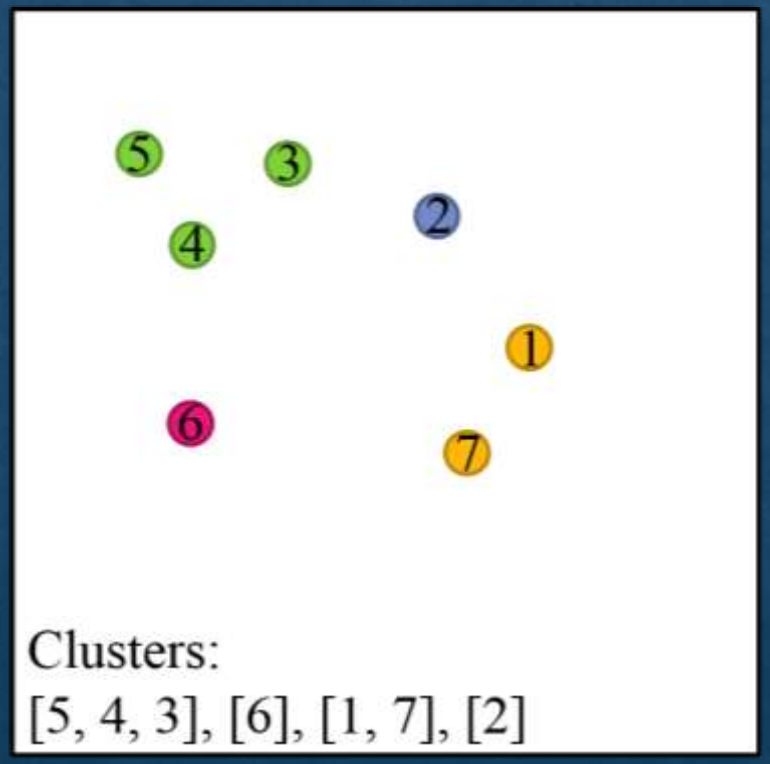
Our Approach – Linking

Our Approach – Linking

- Max 4 indications
  - $\Delta(A, B)$, $\Delta(B, A)$
  - Horizontal and vertical indication
- Assign probability brackets by how many indications we have
  - (0.0, 0.25], (0.25, 0.5], (0.5, 0.75], (0.75, 1.0]
- Sort brackets according to scoring
  - Scoring: how many standard deviations is the distance below the mean?

UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Clustering



UNIVERSITÉ DE NEUCHÂTEL

Our Approach – Clustering

- Cluster document *A* and document *B* if link between them exists
- Follow transitivity rule
  - Link between *A* and *B*, and link between *B* and *C*, then cluster A, B, C

UNIVERSITÉ DE NEUCHÂTEL

Evaluation

- High precision
- Recall acceptable
- Stable performance
- Low MAP
  - Better than baseline
- Slightly better F-Score

		F-Score	Precision	Recall	MAP
Training	Our Approach	0.8184	0.9859	0.7135	0.1036
	Our Baseline	0.8115	1.0000	0.6971	0.0222
Test	Our Approach	0.8218	0.9816	0.7215	0.0540
	Our Baseline	0.8209	1.0000	0.7106	0.0165

UNIVERSITÉ DE NEUCHÂTEL

Conclusion

- Biased towards many authors and small clusters
- Most frequent terms provide discriminative features
- Manhattan distance is a usable measure
- Good performance for simple, unsupervised approach
- Independent of language or genre

UNIVERSITÉ DE NEUCHÂTEL


Thank you for your attention
Mirco Kocher
University of Neuchâtel, Switzerland