Comparing Topic Model Stability Between Finnish, Swedish, English and French

What

We train and compare topic models in a four-way parallel corpus, with and without lemmatisation, with different corpus sizes.

Why

Topic models are usually taken as granted. We aim to determine whether LDA behaves the same way in languages of different families, how useful preprocessing is, and how stable the models are when the corpora are reduced in size.

How

Each model is manually annotated and labels are compared between languages. Same-language models are compared automatically using Jaccard distances and Shannon-Jensen divergences.

Corpus

The parallel corpus consists of all DBpedia abstracts that exist in all four languages. Short abstracts are two-to-three sentences, effectively removing the "cultural differences" between topics in the different languages.

Stop words are removed. For each language, one version of the corpus is lemmatised and the other is not. Topic models are trained on each corpus, the corpora are reduced in size, topic models are trained, etc.

Results

- Type of text plays a big role across all languages:
  - small vocabulary size makes for non-precise topics
  - small document size creates non-precise clustering

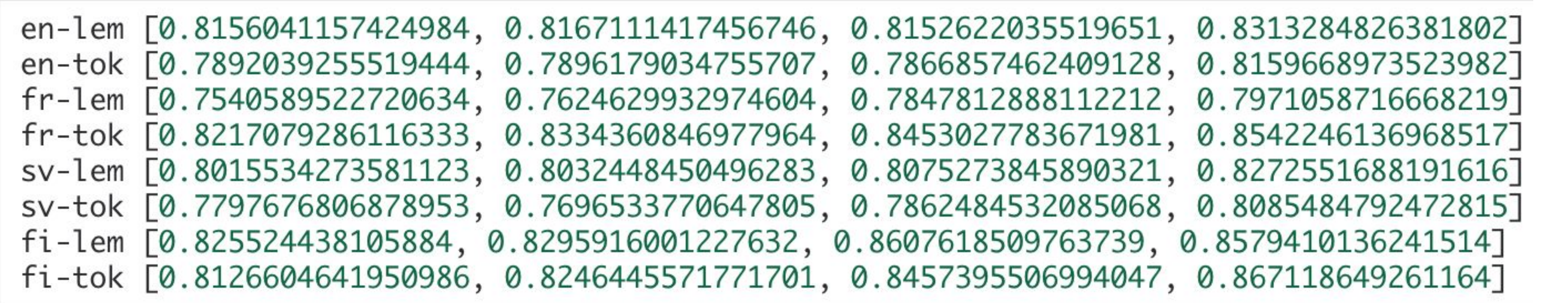
- Unsurprisingly, the above limitations are worsened by corpus size reduction

- Lemmatisation makes labelling harder in English and Swedish, easier in French and Finnish

- Topic models strongly diverge across size, except for a few umbrella topics: "sport", "geography", "music", "biographies", "cinema", "biology". Less strong topics are clustered together.

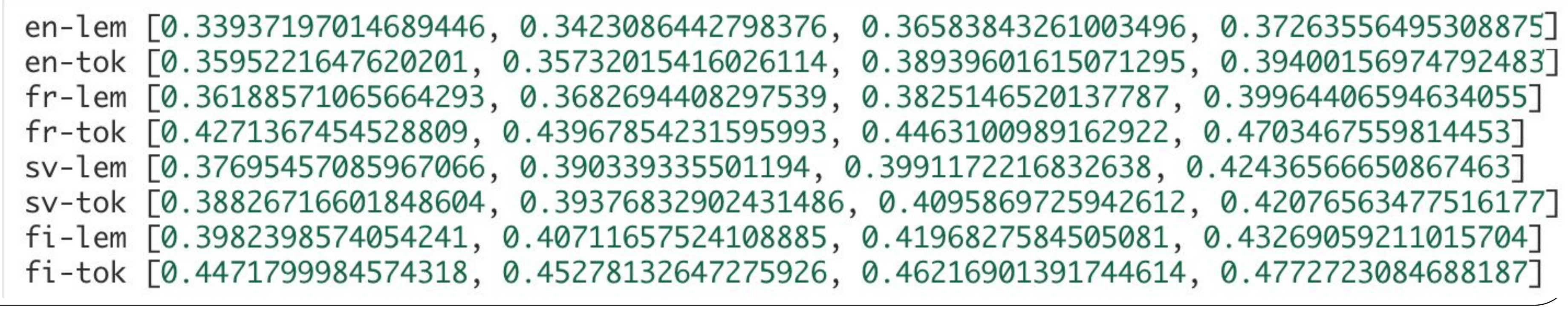
- Some topics appear strongly in all models: Eurovision, ice hockey

Jaccard distances between the full-size corpora and the reduced versions (lower is better)



```
en-lem [0.8156041157424984, 0.8167111417456746, 0.8152622035519651, 0.8313284826381802]
en-tok [0.7892039255519444, 0.7896179034755707, 0.7866857462409128, 0.8159668973523982]
fr-lem [0.7540589522720634, 0.7624629932974604, 0.7847812888112212, 0.7971058716668219]
fr-tok [0.8217079286116333, 0.8334360846977964, 0.8453027783671981, 0.8542246136968517]
sv-lem [0.8015534273581123, 0.8032448450496283, 0.8075273845890321, 0.8272551688191616]
sv-tok [0.7797676806878953, 0.7696533770647805, 0.7862484532085068, 0.8085484792472815]
fi-lem [0.825524438105884, 0.8295916001227632, 0.8607618509763739, 0.8579410136241514]
fi-tok [0.8126604641950986, 0.8246445571771701, 0.8457395506994047, 0.867118649261164]
```

Jensen-Shannon divergences between the full-size corpora and the reduced versions



```
en-lem [0.33937197014689446, 0.3423086442798376, 0.36583843261003496, 0.37263556495308875]
en-tok [0.3595221647620201, 0.35732015416026114, 0.38939601615071295, 0.39400156974792483]
fr-lem [0.36188571065664293, 0.3682694408297539, 0.3825146520137787, 0.39964406594634055]
fr-tok [0.4271367454528809, 0.43967854231595993, 0.4463100989162922, 0.4703467559814453]
sv-lem [0.37695457085967066, 0.390339335501194, 0.3991172216832638, 0.42436566650867463]
sv-tok [0.38826716601848604, 0.39376832902431486, 0.4095869725942612, 0.42076563477516177]
fi-lem [0.3982398574054241, 0.40711657524108885, 0.4196827584505081, 0.43269059211015704]
fi-tok [0.4471799984574318, 0.45278132647275926, 0.46216901391744614, 0.47727230846881 87]
```

References

[1] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. Springer (2007)
[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993{1022 (2003)
[3] Brauer, R., Fridlund, M.: Historicizing topic models, a distant reading of topic modeling texts within historical studies. In: International Conference on Cultural Research in the context of Digital Humanities, St. Petersburg: Russian State Herzen University (2013)
[4] Hengchen, S., O'Connor, A., Munnelly, G., Edmond, J.: Comparing topic model stability across language and size. In: Proceedings of the Japanese Association for Digital Humanities Conference 2016 (2016)
[5] Joachims, T.: Learning to classify text using support vector machines: Methods, theory and algorithms, vol. 186. Kluwer Academic Publishers Norwell (2002)
[6] Jockers, M.L.: Macroanalysis: Digital methods and literary history. University of Illinois Press (2013)
[7] Jockers, M.L., Mimno, D.: Signicant themes in 19th-century literature. Poetics 41(6), 750{769 (2013)
[8] Landauer, T.K., Foltz, P.W., Laham, D.: An introduction to latent semantic analysis. Discourse processes 25(2-3), 259{284 (1998)
[9] May, C., Cotterell, R., Van Durme, B.: Analysis of morphology in topic modeling. arXiv preprint arXiv:1608.03995 (2016)
[10] Munnelly, G., O'Connor, A., Edmond, J., Lawless, S.: Finding meaning in the chaos (2015)
[11] Real, R., Vargas, J.M.: The probabilistic basis of jaccard's index of similarity. Systematic biology 45(3), 380{385 (1996)

comhis.github.io

Simon Hengchen, Antti Kanner, Jani Marjanen and Eetu Mäkelä