ADVANCED NATURAL LANGUAGE PROCESSING

Homework 6

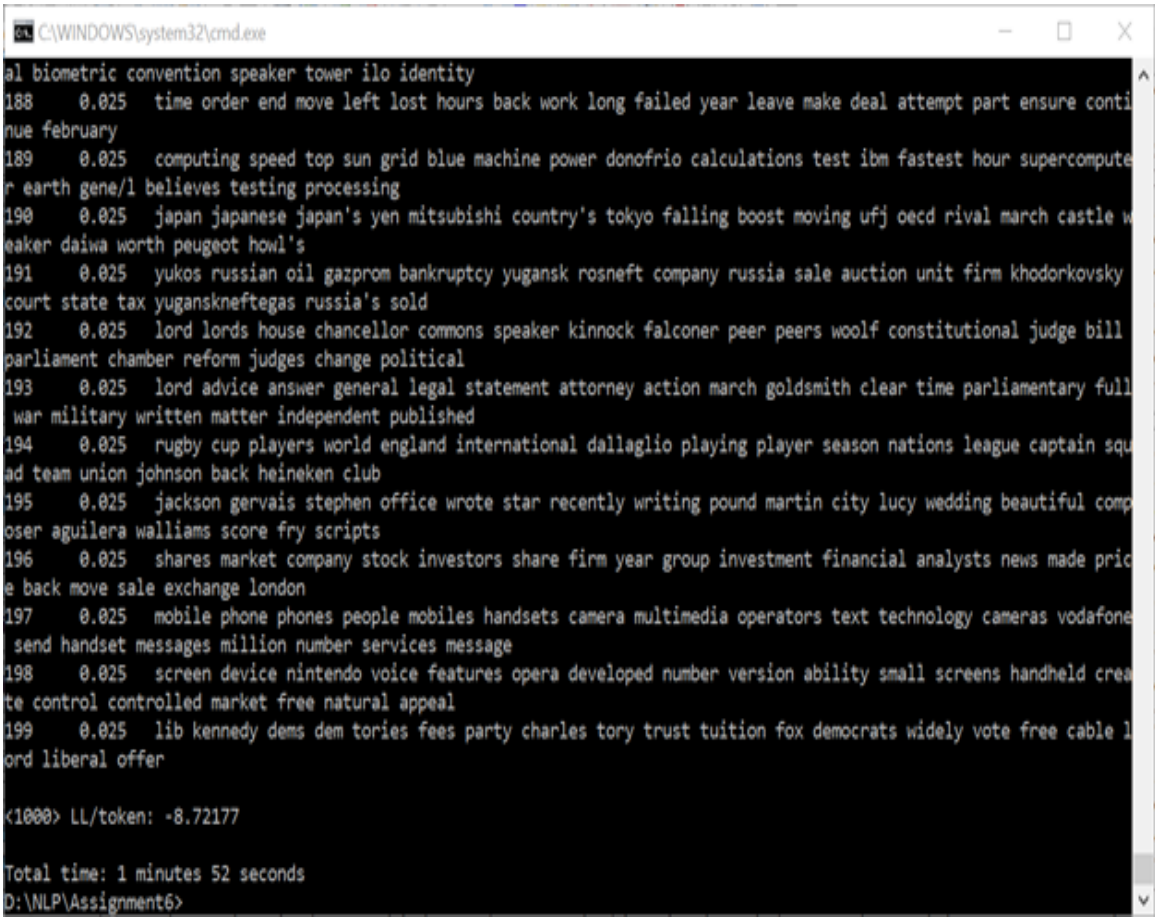
Anusha Ramamurthy

a. Pick some topic modeling software, e.g., one of the following, and get it working:

We initially chose R to do this task. Unfortunately we had trouble installing the library tm and slam.  Mallet seemed to be more user friendly.

b. Pretend you are a social scientist or a humanities scholar and come up with some rough question you want to investigate. Gather a small amount of appropriate data and run the topic modeling software on that data, to learn the models.

We decided to use BBC articles for the year 2004-05. The question we had in mind was to identify the top 200 topics published by BBC for this year. The data was split into different genres, but we decided to run mallet on the entire set. The below is a screenshot of the keywords identified for each topic.



```
al biometric convention speaker tower ilo identity
188     0.025   time order end move left lost hours back work long failed year leave make deal attempt part ensure conti
nue february
189     0.025   computing speed top sun grid blue machine power donofrio calculations test ibm fastest hour supercompute
r earth gene/l believes testing processing
190     0.025   japan japanese japan's yen mitsubishi country's tokyo falling boost moving ufj oecd rival march castle w
eaker daiwa worth peugeot howl's
191     0.025   yukos russian oil gazprom bankruptcy yugansk rosneft company russia sale auction unit firm khodorkovsky
court state tax yuganskneftegas russia's sold
192     0.025   lord lords house chancellor commons speaker kinnock falconer peer peers woolf constitutional judge bill
parliament chamber reform judges change political
193     0.025   lord advice answer general legal statement attorney action march goldsmith clear time parliamentary full
 war military written matter independent published
194     0.025   rugby cup players world england international dallaglio playing player season nations league captain squ
ad team union johnson back heineken club
195     0.025   jackson gervais stephen office wrote star recently writing pound martin city lucy wedding beautiful comp
oser aguilera walliams score fry scripts
196     0.025   shares market company stock investors share firm year group investment financial analysts news made pric
e back move sale exchange london
197     0.025   mobile phone phones people mobiles handsets camera multimedia operators text technology cameras vodafone
 send handset messages million number services message
198     0.025   screen device nintendo voice features opera developed number version ability small screens handheld crea
te control controlled market free natural appeal
199     0.025   lib kennedy dems dem tories fees party charles tory trust tuition fox democrats widely vote free cable l
ord liberal offer

<1000> LL/token: -8.72177

Total time: 1 minutes 52 seconds
D:\NLP\Assignment6>
```

c. Investigate the output of the tool and give a brief assessment. Do the topics (i.e., the most frequent words) make sense? Can you see the documents annotated in a way that helps answer your question? Do any mistakes in the output connect with what you know about how the system is working? If you were a non-computer-savvy social scientist, what would have been the most significant barriers to using the software (if any)?

We chose the command line argument option to save the outputs in file format. The bbc_200.txt gives us the keywords associated with different the top 200 topics.We found 194 occurances of "football" , 402 occurances of America, 334 occurances of India and 17 occurances of "racial" in these 200 topics.

It definitely is a good way to identify the top trends of the year. It's also in my opinion, a good way to identify the bias if there are any. As "church" appears 48 times among 200 topics. If we ran Mallet over a national daily like an Indian Newspaper we would probably have more news about topics around say "temples"

The most glaring mistake is that these words may appear many times in the same article/topic, but one would assume that its the most spoken about word. The mallet picks words at random and the results are based on the order. If we jumble up the words we will not get the same results. If an article uses a lot of matephors and similis to explain the topic, in such cases Mallet would not be able to identify this. For example if he uses dog as an example to illustrate loyality, Mallet might assume that dog is one of the topics. So it definitely depends on the writing style used.

The command line would definitely be intimidating to a non computer savvy person. Also the outputs are not very clear. Understanding and evaluating the results/outputs produced needs more study.

Acknowledgements
Worked in collaboration with Anup Bharadwaj.

References:
http://mlg.ucd.ie/datasets/bbc.html
https://cran.r-project.org/web/packages/topicmodels/index.html
(Links to an external site.)

Citation:
D. Greene and P. Cunningham. "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.