Lab 8: Solutions

Author: Luke Shrimpton
Author: Sharon Goldwater
Date: 2014-11-01, 2015-11-10

Examining the data files

- What word's cooccurrence statistics are listed on the second line of the `counts` file?

  `less /afs/inf.ed.ac.uk/group/teaching/anlp/counts` shows that the `counts` file starts like this:

  ```
  138489679
  0       328506  7550202 13      6270 17 23198 19        19541 18 ...
  ```

  The second line lists the cooccurrence statistics for the word whose word ID is `0`. To figure out which word that is:

  ```
  grep '^0' /afs/inf.ed.ac.uk/group/teaching/anlp/wid_word
  ```

  which shows that the word is `blog`.

- How many distinct words are left in our data set?

  ```
  wc /afs/inf.ed.ac.uk/group/teaching/anlp/wid_word
  ```

  shows that there are 210516 lines in the file, i.e. 210516 distinct words.

Examining and running the code

- How big is the `wid_word` dictionary?

  ```
  len(wid_word)
  ```

- What is the id of `'@bbcnews'` and what word has the id `200`?

  ```
  word_wid['@bbcnews'] wid_word[200]
  ```

  shows id is 14379 and word is `princess`.

- Check whether the following words exist in the dataset: `'bob'`,`'toast'`,`'insektors'`,`'looovvee'`, `'you'`,`'Norman'`,`':-)'`.

  You can check this by doing `word_wid[w]` for each word `w` in the list. You will get a key error for the words that don't exist.

[1] http://creativecommons.org/licenses/by-nc/4.0/.

[2] http://www.inf.ed.ac.uk/teaching/courses/anlp/labs/lab8-sol.py

'bob': exists
'toast': exists
'Norman': doesn't exist because it contains a capital letter. The lowerc ase version norman is in the dataset.
'insektors': doesn't exist, presumably because of low frequency
'looovvee': exists
'you': doesn't exist. This is a high-frequency words so presumably filtered out as a stopword
':-)': doesn't exist because it contains punctuation

- What data type is o_counts[word_wid['love']] and co_counts[word_wid['love']]?
  The first is an integer, the second a dict (keys are word id's, values are counts)
- How many times does the word '@justinbieber' occur in the dataset?
  o_counts[word_wid['@justinbieber']]
  yields 703307.
- How many times does the word '@justinbieber' co-occur with 'love'?
  co_counts[word_wid['@justinbieber']][word_wid['love']]
  yields 120573
- How many times does the word '@justinbieber' co-occur with 'hate'?
  co_counts[word_wid['@justinbieber']][word_wid['hate']]
  yields 4393

Do Twitter users like Justin Bieber?

See lab8-sol.py[2] for code. The PMI values for 'love' and 'hate' are 2.197 and -0.639 respectively, indicating that Twitter users like Justin Bieber. The trend holds up if you add other pos/neg words as well, though exact numbers will vary.

Husbands and wives

- Which of these two words occurs more in this dataset? By how much? What are some possible explanations for this difference? (There are quite a few!)
  husband occurs only about half as much as wife (71258 vs 137077). I originally thought of two reasons: (1) people don't talk about husbands as much as they talk about wives, or (2) there are more men on Twitter than women, and everyone talks about their spouses equally often. Previous students thought of lots of other possibilities:
  - husband/wife are used self-referentially and there are more *women* than *men* on twitter, or women just refer to their role as wife more.
  - women tend to use other words for their husband (eg. hubby, or the husband's name).
  - men just tweet more than women (even if the number of users is similar).
  - Some popular movie, TV show or book with wife in the title came out at the time this data was collected.

  If you were using this dataset for research and the relative counts of these words mattered, you might want to do some further work to figure out the reason for it. It's always good to think about *why* the data is the way it is, don't just take it at face value.

- Are there noticeable differences in the sentiment of tweets in which people refer to husbands or wives? If so, what are they?

  The differences in sentiment are not large. In both cases people are more likely than chance to `love` a spouse (slighly more so for husbands) and less likely than chance to `hate` one (slightly more so for wives).

- What about other family members?

  For `son` and `daughter`, there is a much bigger difference: daughters appear to be more loved.

  People seem to be rather self-hating on Twitter: PMI(*hate*, *self*) is more than 2, whereas PMI(*love*, *self*) is close to 0. Though, if you look at the raw tweets you'll notice that not all instances of `self` refer to the person posting. So, there might be more interesting things going on here. You could try looking at a more specific word like `myself` but you'd need to change the preprocessing because that word has been filtered out of our dataset.

  `child` has small positive sentiment, whereas `kid` has both positive and (more) negative one, indicating very different usage.

  If you expand the list of sentiment words, you're likely to find similar results to the above but less pronounced. You may also find that some of the negative PMI scores are now positive. This seems to indicate that although people might not use the word `hate` specifically, they do still express negative sentiment about the target words.