Course title: Digital Humanities Prof. J. Savoy

Serie 7. Language models and Twitter

In a small experiment, Mary had generated a language model with a training corpus. In this corpus, she counted 555 distinct word-types, and 5,050 tokens. In addition, she obtained the following information.

Unigram	Frequency	Bigram	Frequency
Δ	1000	Δ today	3
the	1350	today the	4
big	320	the big	210
increase	10	big deal	11
deal	15	deal Δ	1
today	25	Δ the	580
bank	8	the stock	8
stock	25	stock is	2
will	56		
is	132		

Now she wants to evaluate the probability of two sequences:

"Today, the big deal" and "the stock is decreasing"

Mary is applying a direct estimation as a probability estimate (also called the maximum likelihood principle). Ann suggests that she can apply Laplace smoothing for better estimations.

Question 1. Compute the probabilities of the two sequences with and without Laplace smoothing.

Question 2. Give one problem when applying the direct estimation as suggested by Mary. Provide two drawbacks related to Laplace smoothing.

In the folder "Corpus", you can find the file `tweets.female.txt` and `tweets.male.txt` containing tweets sent by women or men (in UK). These two files form the training sample (with, of course, the two target categories female and male). The final task would be to classify new tweets according to author gender. The second pair of files (namely `tweet.female.test.txt`, `tweets.male.test.txt`) forms the test sample. Don't use them. For the moment, you can only use the training sample.

The structure of these four files is the following:

userID	'human'	'F/M'	text of the tweet
0	human	F	alex is too nice for love island :(
0	human	F	@ lipstaco @ jennyhastie

The preprocessing was done. The text is already in lowercase, with a space between every token. No stemmer and stopword list were applied. A user is writing usually 100 tweets (for some users, it is a little bit less). All the tweets written by a given user can be identified with the userID (first column).

To read the text of the tweet you can write something like:

```
aLine = myInputFile.readline()
while aLine:
   myLine = aLine.split('\t')
   aFinalLine = ''
   for aSubLine in myLine[3:]:
      aSubLine = re.sub('\n', '', aSubLine)
      aFinalLine += aSubLine + ' '
   aListOfTweets.append(aFinalLine)
   aLine = myInputFile.readline()
```

There is no guarantee that the `split()` function will return always exactly four parts. It could be more than four (e.g., when a comma is present in a tweet).

Question 3. Generate the bigrams of tokens. Provide the ten most frequent bigrams from tweets written by a woman or a man. When looking at these two lists of bigrams, can you infer why they present differences, or can you see a justification for these possible differences between these lists.

On the other hand, one can assume that the author gender can't explain these differences because they are due to various random factors (e.g., as the text is written in (UK) English, I can expect as frequent bigrams seeing 'of the', 'thank u' or 'to be').