Text Mining Exercise week 6 report

Guido Zuidhof, s4160703

Reference: `At least 13 sailors have been killed in a mine attack on a convoy in north-western Sri Lanka officials say` (20 unigrams)

system-1: `Tamil Tiger guerrillas have blown up a navy bus in northeastern Sri Lanka killing at least 10 sailors and wounding 17 others` (22 unigrams)

system-2: `Blasts blamed on Tamil Tiger rebels killed 13 people in Wednesday in Sri Lanka's northeast and dozens more were injured officials said, raising fears planned peace talks may be cancelled and a civil war could restart` (37 unigrams)

***1.)* Explain the Log-likelihood ratio test for topic signatures in your own words using a simple example**

The Log-likelihood ratio test can be used to determine how "identifying" a piece of text (usually a sentence) is. If it is not used often in a big dataset (your prior available corpus for instance), but it does occur often in a certain piece of text, it may be more likely that it is important and should thus be included in a summarization.

The importance of sentences could be quantified by the likelihood of the words in the sentence. Every word could be assigned a weight of how likely it is (the topic signature).

If in my large corpus the word `fruit` occurs very often, it would have a different weight from the word `advocado` if that occurs less often. If in a sentence the word `advocado` occurs very often, it may be very important, and get a high *importance* score, as opposed to a sentence containing the word `fruit` just as many times.

***2a.)* Compute recall, precision and F-score on the unigrams for system 1 and system 2.**

systen-1

```
Recall: 7 / (7 + 11) = 0.389
Precision: 7 / (7 + 15) = 0.318
F-score: (0.389 * 0.318) / (0.389 + 0.318) = 0.175
```

system-2

```
Recall: 8 / (8 + 10) = 0.444
Precision: 8 / (8 + 27) = 0.229
F-score: (0.444 * 0.229) / (0.444 + 0.299) = 0.137
```

***2b.)* Compute ROUGE-1 and ROUGE-L (longest common subsequence) for both systems. Include the LCS that you found in your answer.**

Amount of unigrams in reference: `18`.
Amount of matches for system-1: `7`.
Amount of matches for system-2: `8`.

System-1 `ROUGE-1 = 7 / 18 = 0.389`.
System-2 `ROUGE-1 = 8 / 18 = 0.444`.

```
System-1 and reference LCS: Sri Lanka
LCS length: 2
ROUGE-L = 2 / 18 = 0.111
```

```
System-2 and reference LCS: Sri Lanka
LCS length: 2
ROUGE-L = 2 / 18 = 0.111
```