Information Retrieval Exercises
Assignment 5:
Collocations

Samuele Garda (gardasam@informatik.hu-berlin.de)
Collocations

- A **collocation** is a series of words or terms that co-occur more often than would be expected by chance (Wikipedia).

- Two terms co-occur if they appear together in a context (e.g. a sentence or a window of $n$ words).

- If two words are independently very frequent, their co-occurrence is less statistically significant than if the two words are comparatively rarer:
  - “Peace Corps” is more interesting than “are going”
Finding Collocations

• Find top ranked collocations in IMDB corpus

• Parse the title and plot descriptions from the plot.list

• Use pre-processing from assignment 2
  – Tokenization at spaces, line breaks, dots, commas, colons, question marks and exclamation marks ([ .,:!?])
  – Lower-case all tokens

• Since we don’t detect sentence borders, we only consider subsequent occurrences of the two tokens as co-occurrence!
Finding Collocations

• Consider only bigrams:
  - tokens have to be **adjacent** to one another in the corpus

• Disregard all tokens that are **stop words**:
  - According to this list: https://www.ranks.nl/stopwords
  - Don’t remove stop words from the corpus, just disregard collocations containing them

• Disregard infrequent tokens, i.e.: **less than 1000 total occurrences in the corpus**
Finding Frequent Co-occurrences

• Sort collocations in descending order according to this **association measure**:

\[
s(t, t') = \frac{2 \cdot F(t, t')}{F(t) + F(t')}
\]

- \(F(t, t')\): is the frequency of bigram \((t, t')\) in the corpus
- \(F(t)\): is the frequency of token \(t\) in the corpus

• Return only the top 1000 co-occurrences and their score
Example

• Sentences
  1) the crystal clear water rose against the coast, merging with the sky
  2) let me be crystal clear about this, Rose
  3) the red sun **rose and the sky** turned clear

• Frequencies:
  - **Tokens**: \( F(\text{crystal}) = 2, \ F(\text{clear}) = 3, \ F(\text{water}) = 1, \ F(\text{rose}) = 3, \ F(\text{sky}) = 2, \ldots \)
  - **Bigrams**: \( F(\text{crystal,clear}) = 2, \ F(\text{water,rose}) = 1, \ F(\text{rose,sky}) = 0, \ldots \)
Example

- Co-occurrence scores:

\[ s(\text{crystal, clear}) = \frac{2 \cdot F(\text{crystal, clear})}{F(\text{crystal}) + F(\text{clear})} = \frac{2 \cdot 2}{2 + 3} = \frac{4}{5} \]

\[ s(\text{water, rose}) = \frac{2 \cdot F(\text{water, rose})}{F(\text{water}) + F(\text{rose})} = \frac{2 \cdot 1}{1 + 3} = \frac{1}{2} \]

\[ s(\text{rose, sky}) = \frac{2 \cdot F(\text{rose, sky})}{F(\text{rose}) + F(\text{sky})} = \frac{2 \cdot 0}{3 + 2} = 0 \]
Computation details

• Title, plot and plots from different authors are **treated as as different texts**

| MV: “The Simpsons” (2018) {Springfield Splendor} |
| PL: The best |
| PL: episode. |

BY: foo@example.com

| PL: A rather dull episode. |

| **Potential bigrams:** |
| the simpsons |
| the best |
| best episode |
| a rather |
| rather dull |
| dull episode |

• **Note:** Some of these bigrams will later be discarded due to containing a stop word or infrequent word!
Submission

• No Java class skeleton given this time
  – You may reuse your code from the other assignments!
• Submit executable JAR CoOccurrencesFinder.jar
  – Syntax: java –jar CoOccurrencesFinder.jar <plot-file> <output-file>

• Write top 1000 co-occurrences sorted (desc) by score to <output-file>
  – Syntax: <token>	<token>	<score>
  – los angeles 0.8932607215793057
  – hong kong 0.7493632195618951
  – las vegas 0.7398075240594926
  – us 0.70640263377721
  – united states 0.6942972495584153
Competition

• As fast as possible:
  – Parse corpus
  – Collect co-occurrences

• Use memory abundantly (you have up to 50 GB)
Checklist

• Before submitting your results, make sure that you:

  − ... named your jar CoOccurrencesFinder.jar

  − ... included your source code in the submitted archive

  − ... tested your executable JAR on gruenau:
    java -jar CoOccurrencesFinder.jar plot.list output.txt
    (you might have to increase Java heap space, e.g. -Xmx6g)

  − ...made sure the output is syntactically correct
Roadmap for the last weeks

- **06./07.07.2021**
  - Evaluation and presentation of assignment 4 solutions
  - Q/A for assignment 5

- **09.07.2021, 23:59 (midnight)**
  - Submission deadline for assignment 5 (all groups)

- **13./14.07.2021**
  - Evaluation and presentation of assignment 5 solutions
  - Feedback & farewell
Questions?