Human-centered Artificial Intelligence

2024/2025

Worksheet:
Text Mining

Hugo Oliveira, Luís Macedo

1.1 Topics

- Text Mining
- Sentiment Analysis / Opinion Mining

1.2 Pre-class Materials

- Course slides on Text Mining
- Course slides on Sentiment Analysis
- Video on Sentiment Analysis:
  https://www.youtube.com/watch?v=n4L5hHFcGVk (13min)
- Video *Word Embeddings - EXPLAINED!*:
  https://www.youtube.com/watch?v=GmXkCCa4eVA (9min)

1.3 Complementary Materials

- Course slides on Data Collection & Annotation
- Jurafsky and Martin [2009] (Chapters 4, 6, 9, 10, 11)
- Aggarwal [2015] (Chapter 13)
- Bing Liu's lecture on Sentiment Analysis (slides http://www.cs.uic.edu/~liub/FBS/sentiment-analysis-tutorial-2012.pdf, video https://www.youtube.com/watch?v=6ki_F_5a5vg)
- Video *What Are Word and Sentence Embeddings?*:
  https://www.youtube.com/watch?v=A8HEPBdKVMA (8min)
- Video *Text Representation using Word Embeddings*:
  https://www.youtube.com/watch?v=Do8cVbx-HOs

1.4 Theoretical-Practical Exercises

Question 1.1 What Text Mining techniques would be useful for the following scenarios?

- You were provided with several gigabytes of textual documents, of which you know nothing about, but believe that they might include valuable explanations for some mysteries of human history. To help you explore, you want to search those documents for specific keywords or phrases.
- You want to have a high-level picture of the most common themes in the previous collection, without having to read through all the documents.
- You note that about half of the documents have a naming convention that might be related to the actual contents of the document. Towards a better organisation, you want to predict the names of the files that do not match this naming convention.
- You want to develop a tool for helping in the process of writing emails, i.e., for every received email, it should propose a possible response.

Question 1.2 Consider the following text corpus with six documents:

D	Content
d1	`Good-looking food... that tasted bad!!`
d2	`Service not good. Food not good.`
d3	`Terrible, terrible food.`
d4	`Amazing tasting food. Best value.`
d5	`Good service. Good food. Good value for money.`
d6	`Look: best-tasting food.`

Now answer the following questions:

a. If you were doing Text Mining from documents of this kind, would you perform any kind of pre-processing?

b. Adopt the term-document matrix, based on frequency counts, for representing the term-document matrix of this corpus. In the process, ignore punctuation signs, prepositions (that, for) and lemmatise verbs in the gerund (*looking*, *tasting*).

c. Using the previous vector representation, compute the similarity between: d1 and d3; d1 and d4; d1 and d5.

d. Which tasks in the domain of Sentiment Analysis could be applied to this data? Clarify their goal and the result for each document.

e. What would be the limitations of relying exclusively on a vector representation based on frequency counts, and possibly a sentiment lexicon? How could they be minimised, for different Sentiment Analysis tasks? Consider different scenarios regarding annotated data and computer power available.

f. Consider that the polarity of these documents was annotated independently, by two humans, who tagged them as follows:

- Ana: Neutral, Neutral, Negative, Positive, Positive, Positive
- Hugo: Negative, Negative, Negative, Positive, Positive, Positive

How would you classify the annotator agreement?

Question 1.3 Consider the following corpus of training data and answer the questions:

```
<s> I am Sam </s>
<s> Sam I am </s>
<s> Sam I like </s>
<s> Sam I do like </s>
<s> do I like Sam </s>
```

a. With a unigram, bigram and trigram model...

1. What is the most probable word predicted for the following sequences?
   - `<s> Sam ...`
   - `<s> Sam I do ...`
   - `<s> Sam I am Sam ...`
   - `<s> do I like ...`
2. Which of the following sentences gets a higher probability?
   - `<s> Sam I do I like </s>`
   - `<s> Sam I am </s>`

- `<s> Sam I do like Sam I am </s>`

b. Discuss the limitations of N-gram language models and give examples.

Bibliography

Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Prentice Hall series in Artificial Intelligence. Prentice Hall, Pearson Education International, Englewood Cliffs, NJ, 2nd edition, 2009.