NOVA IMS Information Management School

Text Mining | Spring 2022

Homework #1

Deadline: May 20th, 11:59 PM (aka 23h59)

Please answer to each question in a complete way, showing the relevant intermediate steps to your results (when applicable). Good work!

1. For each of the regular expressions below, select all the options that would be completely recognized by each of those expressions (if any):

   1.1. **(A|B)BB?[AEIOU]**

   a) ABBA
   b) BBBB
   c) AI
   d) BAI
   e) None of the above
   f) All of the above

   1.2. **CO+L?[AO]**

   a) COA
   b) COOL
   c) COLA
   d) CAO
   e) None of the above
   f) All of the above

2. Write a regular expression that recognizes e-mails in the format <username>@<domain>, where <username> may only contain one or more letters, digits, underscores (_), or dots, and <domain> may only contain one or more lowercase letters, followed by either ".com", ".net", or ".org".

3. Consider the following (already preprocessed) sentences:

   a) i really like soccer
   b) i like basketball
   c) soccer really is my favorite sport

   3.1. Which sentence (**b** or **c**) is the most similar to **a**) according to **Minimal Edit Distance (MED)**? Fill and show the matrices to obtain **MED(a,b)** and **MED(a,c)**, and indicate which edit operations would need to occur according to each of the matrices obtained. Consider a word-level comparison and a cost of 1 for each edit operation.

   3.2. Which sentence (**b** or **c**) is the most similar to **a**) according to **Dice**? Obtain **Dice(a,b)** and **Dice(a,c)**. Consider a word-level comparison, and round the results to two decimal places.

4. Consider a collection **D** containing the following documents:

   a) The Special One will make changes to the starting eleven
   b) Eleven has special powers
   c) Top 11 most special games

   4.1. Rewrite the documents above after applying the following preprocessing:

   - Lowercasing
   - Removing the following stop-words: **the**, **will**, **to**, **has**
   - Replacing every sequence of **digits** by the token **[DIGIT]**

   4.2. Obtain the **bag-of-words vectors** (binary vectors) for the documents above, considering the preprocessing you did in **4.1.**

   4.3. Which document (**b** or **c**) is the most similar to **a**) according to the **cosine similarity** of the bag-of-words vectors? Obtain **cos_sim(a,b)** and **cos_sim(a,c)**, rounding the results to two decimal places.

   4.4. Compute the **TD-IDF** of the terms **“special”** and **“eleven”** with respect to the document **a**) and the collection of documents **D**. Which of these two terms is the most useful for discriminating documents in this collection?