Introduction to Web Science

Assignment 8

Prof. Dr. Steffen Staab
staab@uni-koblenz.de

René Pickhardt
rpickhardt@uni-koblenz.de

Korok Sengupta
koroksengupta@uni-koblenz.de

Olga Zagovora
zagovora@uni-koblenz.de

Institute of Web Science and Technologies
Department of Computer Science
University of Koblenz-Landau

Submission until: January 11, 2017, 10:00 a.m.
Tutorial on: January 13, 2017, 12:00 p.m.

Please look at all the lessons of part 2 in particular **Similarity of Text** and **graph based models**

For all the assignment questions that require you to write code, make sure to include the code in the answer sheet, along with a separate python file. Where screen shots are required, please add them in the answers directly and not as separate files.

Other than that this sheet is mainly designed to review and apply what you have learnt in part 2 it is a little bit larger but there is also more time over the x-mas break. In any case we wish you a mery x-mas and a happy new year.

Team Name: XXXX

1 Similarity - (40 Points)

This assignment will have one exercise which is dived into four subparts. The main idea is to study once again the web crawl of the Simple English Wikipedia. The goal is also to review and aply your knowledge from part 2 of this course.

We have constructed two data sets from it which are all the articles and the link graph extracted from Simple English Wikipedia. The extracted data sets are stored in the file `http://141.26.208.82/store.zip` which contains a pandas container and can be read with pandas in python. In subsection "1.5 Hints" you will find some sample python code that demonstrates how to easily access the data.

With this data set you will create three different models with different similarity measures and finally try to evaluate how similar these models are.

This assignment requires you to handle your data in efficient data structures otherwise you might discover runtime issues. So please read and understand the full assignment sheet with all the tasks that are required before you start implementing some of the tasks.

1.1 Similarity of Text documents (10 Points)

1.1.1 Jaccard - Similarity on sets

1. Build the word sets of each article for each article id.
2. Implement a function `calcJaccardSimilarity(wordset1, wordset2)` that can calculate the jaccard coefficent of two word sets and return the value.
3. Compute the result for the articles `Germany` and `Europe`.

1.1.2 TF-IDF with cosine similarity

1. Count the term frequency of each term for each article
2. Count the document frequencies of each term.
3. For each article id provide a dictionary of terms occuring in the article together with their tf-idf scores as the corresponding values.
4. Implement a function `calculateCosineSimilarity(tfIdfDict1, tfIdfDict2)` that computes the cosine similarity for two sparse tf-idf vectors and returns the value.
5. Compute the result for the articles `Germany` and `Europe`.

1.2 Similarity of Graphs (10 Points)

You can understand the similarity of two articles by comparing their sets of outlinks (and see how much they have in common). Feel free to reuse the `computeJaccardSimilarity` function from the first part of the exercise. This time do not aply it on the set of words within two articles but rather on the set of outlinks being used within two articles. Again compute the result for the articles `Germany` and `Europe`.

1.3 How similar have our similarities been? (10 Points)

Having implemented these three models and similarity measures (text with Jaccard, text with cosine, graph with Jaccard) our goal is to understand and quantify what is going on if they are used in the wild. Therefore in this and the next subtask we want to try to give an answere to the following questions.

- Will the most similar articles to a certain article always be the same independent which model we use?
- How similar are these measures to each other? How can you statistically compare them?

Assume you could use the similarity measure to compute the top k most similar articles for each article in the document collection. We want to analyze how different the rankings for these various models are.

Do some research to find a statistical measure (either from the lectures of part 2 or by doing a web search and coming up with something that we haven't discussed yet) that could be used best to compare various rankings for the same object.

Explain in a short text which measure you would use in such an experiment and why you think it is usefull for our task.

1.4 Implement the measure and do the experiment (10 Points)

After you came up with a measure you will most likely run into another problem when you plan to do the experiment.

Since runtime is an issue we cannot compute the similarity for all pairs of articles. Tell us:

1. How many similarity computations would have to be done if you wished to do so?
2. How much time would roughly be consumed to do all of these computations?

A better strategy might be to select a couple of articles for which you could compute your meassure. One strategy would be to select the 100 longest articles. Another strategy might be to randomly select 100 articles from our corpus.

Computer your three similarity measures and evaluate them for these two strategies of selecting test data. Present your results. Will the results depend on the method for selecting articles? What are your findings?

Answer: Code:

```
# assignment 8
# Andrea Mildes - mildes@uni-koblenz.de
# Sebastian Blei - sblei@uni-koblenz.de
# Johannes Kirchner - jkirchner@uni-koblenz.de
# Abdul Afghan - abdul.afghan@outlook.de

import logging
import time
import pandas as pd
import re
import numpy as np
from numpy import zeros
import operator
np.set_printoptions(threshold=np.nan)


document_freq = {}
word_vector = {}
article_dic = {}
df1 = pd.DataFrame()
df2 = pd.DataFrame()

# Configure logging and set start time
logging.basicConfig(filename='similarity.log', level=logging.DEBUG)
start_time = 0


# Read the given file into a string
def read_file(file):
    with open(file) as f:
        data = f.read().replace('\n', '')
    return data


# Write a file
def write_file(filename, generated_string):
    with open(filename, 'w') as f:
        f.write(generated_string)

    t = str(round((time.time() - start_time), 2)).zfill(5)
    logging.info("[" + t + "] " +
                 "Finished writing \"" + filename + "\" file. \n\n")
    return

```

```
45:
46: def create_set(s):
47:     regex = re.compile("\w+")
48:     s_list = regex.findall(s)
49:     word_set = set()
50:
51:     for s in s_list:
52:         word_set.add(s.lower())
53:
54:     return word_set
55:
56:
57: def calc_term_freq(s):
58:     regex = re.compile("\w+")
59:     s_list = regex.findall(s)
60:     word_rank = {}
61:
62:     for s in s_list:
63:         s = s.lower()
64:         try:
65:             # Increase count by one
66:             word_rank[s] += 1
67:         except KeyError:
68:             # Create new dictionary entry for terms that
69:             # are not part of the dict yet
70:             word_rank[s] = 1
71:
72:     return word_rank
73:
74:
75: def calc_document_freq(s):
76:     # create set of the given string in order to get only one
77:     # occurrence of each word
78:
79:     s_set = create_set(s)
80:
81:     for s in s_set:
82:         try:
83:             # Increase count by one
84:             document_freq[s] += 1
85:         except KeyError:
86:             # Create new dictionary entry for chars that
87:             # are not part of the dict. yet
88:             document_freq[s] = 1
89:     return
90:
91:
92: def calcJaccardSimilarity(wordset1, wordset2):
93:     intersection = wordset1.intersection(wordset2)
```

```
94:     union = wordset1.union(wordset2)
95:
96:     jac = len(intersection) / len(union)
97:     return jac
98:
99:
100: # Compute the document frequency and store it in global
101: # dictionary document_freq
102: def log_doc_freq():
103:     t = str(round((time.time() - start_time), 2))
104:     logging.info("[" + t + "]  Compute document frequency ...")
105:
106:     df1.text.apply(calc_document_freq)
107:
108:     # sorted_x = sorted(document_freq.items(), key=operator.itemgetter(1))
109:     # for k, v in sorted_x:
110:     #     print(str(v) + " | " + k)
111:
112:     t = str(round((time.time() - start_time), 2))
113:     logging.info("[" + t + "]  Done! \n\n")
114:
115:     return
116:
117:
118: # tfidf(word, document) = tf(word, document) * log(|D| / df(word))
119: def calc_tfidf(term, term_freq):
120:     # 1: term frequency of the word in document
121:     # 2: Amount of Documents
122:     amount_of_documents = len(df1)
123:
124:     if len(document_freq) == 0:
125:         logging.error("document_freq dictionary is empty!")
126:         exit()
127:
128:     # 3: document frequency of the term
129:     document_freq_word = document_freq[term]
130:
131:     tfidf = term_freq * np.log(amount_of_documents / document_freq_word)
132:
133:     return tfidf
134:
135:
136: def create_dic_of_all_article_with_terms():
137:     t = str(round((time.time() - start_time), 2))
138:     logging.info("[" + t + "]  Compute tfidf dictionary ...")
139:
140:     dic = {}
141:
142:     # Iterate over df1
```

```
143:     for i in range(0, len(df1)):
144:         articles_dic = {}
145:         word_rank_dic = calc_term_freq(df1.loc[i].text)
146:
147:         # Iterate over the word rank dictionary of the article df1.loc[i]
148:         for k, v in word_rank_dic.items():
149:             # Store each word as key in the article_dic and the
150:             # tfidf of the word as value
151:             articles_dic[k] = calc_tfidf(k, v)
152:
153:         # Store the resulting article_dic as a value and the
154:         # article id as a key in the dic dictionary
155:         dic[i] = articles_dic
156:
157:     t = str(round((time.time() - start_time), 2))
158:     logging.info("[" + t + "]  Done! \n\n")
159:
160:     return dic
161:
162:
163: # Generate a dictionary with a unique vector for each word
164: def create_word_vectors():
165:     global word_vector
166:
167:     t = str(round((time.time() - start_time), 2))
168:     logging.info("[" + t + "]  Compute word vectors ...")
169:
170:     # Get the amount of all words in all documents
171:     l = len(document_freq)
172:     i = 0
173:
174:     for k, v in document_freq.items():
175:         v = zeros(l)
176:         v[i] = 1
177:         word_vector[k] = v
178:         i += 1
179:
180:     t = str(round((time.time() - start_time), 2))
181:     logging.info("[" + t + "]  Done! \n\n")
182:
183:     return
184:
185:
186: def calculateCosineSimilarity(tfIdfDict1, tfIdfDict2):
187:     # Calculate the vector for tfIdfDict1
188:     article_vector1 = np.zeros(len(document_freq))
189:     for term, tfidf in tfIdfDict1.items():
190:         # Get the vector corresponding to the current word
191:         vec = word_vector[term]
```

```
192:         # Multiply vector with tfidf
193:         article_vector1 += vec * tfidf
194:
195:     # Calculate the vector for tfIdfDict2
196:     article_vector2 = np.zeros(len(document_freq))
197:     for term, tfidf in tfIdfDict2.items():
198:         # Get the vector corresponding to the current word
199:         vec = word_vector[term]
200:         # Multiply vector with tfidf
201:         article_vector2 += vec * tfidf
202:
203:     # Get the dot product of the vectors
204:     dot = article_vector1.dot(article_vector2)
205:
206:     # Get the length of both vectors
207:     length_vector1 = np.linalg.norm(article_vector1)
208:     length_vector2 = np.linalg.norm(article_vector2)
209:
210:     # Inverse cosine of the dot product divided by the product of
211:     # the length of both vectors
212:     cosine_sim = dot / (length_vector1*length_vector2)
213:
214:     return cosine_sim
215:
216:
217: # Calculate the length of all articles and choose the longest ones
218: def get_longest_articles():
219:     l = []
220:
221:     # Iterate over df1 and append a tuple to the list l
222:     # The tuple contains the index of the text combined with its length
223:     for i in range(0, len(df1)):
224:         l.append((i, len(df1.loc[i].text)))
225:
226:     # Sort the list by text-length descending in place
227:     l.sort(key=lambda tup: tup[1], reverse=True)
228:
229:     # Generate list with 100 entries containing only the article ids
230:     l2 = []
231:     for i in range(0, 100):
232:         l2.append(l[i][0])
233:
234:     return l2
235:
236:
237: # Select 100 random articles
238: def get_random_articles():
239:     l = []
240:
```

```
241:     for i in range(0, 100):
242:         l.append(np.random.choice(df1.index, replace=False))
243:
244:     return l
245:
246:
247: def compute_jaccard_similarity_of_all_articles(articles):
248:     t = str(round((time.time() - start_time), 2))
249:     logging.info("[" + t + "]  Compute jaccard similarities "
250:                            "of all articles ...")
251:
252:     matrix = np.zeros(shape=(100, 100))
253:
254:     # Compute every possible combination of articles. We only need to
255:     # calculate the similarity once per pair and we do
256:     # not need to calculate the similarity of the article with itself.
257:     for i in range(0, len(articles)):
258:         i_article = create_set(df1.loc[articles[i]].text)
259:
260:         for j in range(0, i):
261:             j_article = create_set(df1.loc[articles[j]].text)
262:
263:             jaccard = calcJaccardSimilarity(i_article, j_article)
264:             matrix[i, j] = jaccard
265:
266:         if i % 10 == 0:
267:             logging.info("i: " + str(i))
268:
269:     t = str(round((time.time() - start_time), 2))
270:     logging.info("[" + t + "]  Done! \n\n")
271:
272:     return matrix
273:
274:
275: def compute_cosine_similarity_of_all_articles(articles):
276:     t = str(round((time.time() - start_time), 2))
277:     logging.info("[" + t + "]  Compute cosine similarities "
278:                            "of all articles ...")
279:
280:     matrix = np.zeros(shape=(100, 100))
281:
282:     # Compute every possible combination of articles. We only need to
283:     # calculate the similarity once per pair and we do
284:     # not need to calculate the similarity of the article with itself.
285:     for i in range(0, len(articles)):
286:         i_article = article_dic[articles[i]]
287:
288:         for j in range(0, i):
289:             j_article = article_dic[articles[j]]
```

```
290:
291:             cos_sim = calculateCosineSimilarity(i_article, j_article)
292:             matrix[i, j] = cos_sim
293:
294:         if i % 10 == 0:
295:             logging.info("i: " + str(i))
296:
297:     t = str(round((time.time() - start_time), 2))
298:     logging.info("[" + t + "]  Done! \n\n")
299:
300:     return matrix
301:
302:
303: def pretty_print_matrix(matrix):
304:     for i in range(0, 100):
305:         for j in range(0, 100):
306:             print("%1.3f|" % (matrix[i, j]), end='')
307:         print()
308:
309:
310: def main():
311:     # Set start time
312:     global start_time, df1, df2, article_dic
313:     start_time = time.time()
314:
315:     # Reset log file
316:     with open('similarity.log', 'w'):
317:         pass
318:
319:     t = str(round((time.time() - start_time), 2))
320:     logging.info("[" + t + "] --- Started --- \n\n")
321:
322:     store = pd.HDFStore('store2.h5')
323:     df1 = store['df1']
324:     df2 = store['df2']
325:
326:     # ------------------------------------------------#
327:     # 1.1.1 Calculate the Jaccard coefficient for the #
328:     # articles "Germany" and "Europe"                 #
329:     # ------------------------------------------------#
330:     word_set_germany = create_set(
331:         str(df1[df1.name == 'Germany'].text.values[0]))
332:     word_set_europe = create_set(
333:         str(df1[df1.name == 'Europe'].text.values[0]))
334:     print(calcJaccardSimilarity(word_set_germany, word_set_europe))
335:
336:     # ------------------------------------------------#
337:     # 1.1.2 Calculate the cosine similarity for the #
338:     # article "Germany" and "Europe"                #
```

```
339:     # ----------------------------------------------#
340:     t = str(round((time.time() - start_time), 2))
341:     logging.info("[" + t + "] --- Started calculations for 1.1.2 --- \n\n")
342:
343:     # Compute the document frequency and store it in
344:     # global dictionary document_freq
345:     log_doc_freq()
346:
347:     # Compute a dictionary with the article id as key and
348:     # a dictionary for each article containing the article terms
349:     # and the corresponding tfidf values as value
350:     article_dic = create_dic_of_all_article_with_terms()
351:
352:     # Compute a dictionary to match a unique vector to a specific word.
353:     # This is done before calculating the cosine similarity
354:     # so this only has to be executed once
355:     create_word_vectors()
356:
357:     # Get ID of the article "Germany" and "Europe"
358:     ger_id = df1[df1.name == "Germany"].index[0]
359:     eur_id = df1[df1.name == "Europe"].index[0]
360:
361:     # Calculate the cosine similarity of both articles
362:     cos_sim = calculateCosineSimilarity(article_dic[ger_id],
363:                                         article_dic[eur_id])
364:     print(cos_sim)
365:
366:     t = str(round((time.time() - start_time), 2))
367:     logging.info("[" + t + "] --- Finished calculations for 1.1.2 --- \n\n")
368:
369:     # --------------------------#
370:     # 1.2 Similarity of graphs  #
371:     # --------------------------#
372:     t = str(round((time.time() - start_time), 2))
373:     logging.info("[" + t + "] --- Started calculations for 1.2 --- \n\n")
374:
375:     germany_outlinks = set(df2[df2.name == "Germany"].out_links.values[0])
376:     europe_outlinks = set(df2[df2.name == "Europe"].out_links.values[0])
377:     outlinks_jaccard_sim =  calcJaccardSimilarity(germany_outlinks,
378:                                                   europe_outlinks)
379:     print(outlinks_jaccard_sim)
380:
381:     t = str(round((time.time() - start_time), 2))
382:     logging.info("[" + t + "] --- Finished calculations for 1.2 --- \n\n")
383:
384:     # --------------------------#
385:     # 1.4 Implement the measure #
386:     # --------------------------#
387:     t = str(round((time.time() - start_time), 2))
```

```
388:     logging.info("[" + t + "] --- Started calculations for 1.4 --- \n\n")
389:
390:     # Find the 100 longest articles
391:     l_articles = get_longest_articles()
392:
393:     jaccard_matrix = compute_jaccard_similarity_of_all_articles(l_articles)
394:     pretty_print_matrix(jaccard_matrix)
395:
396:     cosine_matrix = compute_cosine_similarity_of_all_articles(l_articles)
397:     pretty_print_matrix(cosine_matrix)
398:
399:     # Find 100 random articles
400:     #r_articles = get_random_articles()
401:     #compute_similarity()
402:
403:     t = str(round((time.time() - start_time), 2))
404:     logging.info("[" + t + "] --- Finished calculations for 1.4 --- \n\n")
405:
406:     t = str(round((time.time() - start_time), 2))
407:     logging.info("[" + t + "] --- Finished --- \n\n")
408:
409:
410: if __name__ == '__main__':
411:     main()
```

Word Rank Frequency Diagram:

Answer to 1.3:
The most similar articles to a certain article will be different, depending on which model is used. The reason for this is because the Jaccard Similarity only checks the occurrence of a term in two documents. The amount of occurrences, however, is not being considerd. The Cosine Similarity on the other hand takes these occurrences into account and the similarity rating decreases, the more often the word occurres in a document. This would lead to different similarity rankings, because the Jaccard Similarity would rank documents with many filler terms higher than the Cosine Similarity, where the rating of those filler terms should be quite low due to the common usage.
One could measure the rankings by setting up a query to a certain subject, as for example a famous person in a certain field, and check if information about this person / subject is ranked higher than information not related to the person / subject. As an example, we take a famous musician and set up a query. The top k most similar articles should be about the person himself, his band, if any, other musicians who play the same instrument or other musicians who played with him.

1.5 Hints:

1. In oder to access the data in python, you can use the following pice of code:
2. Variables df1 and df2 are pandas DataFrames which is tabular data structure. df1 consists of article's texts, df2 represents links from Simple English Wikipedia articles. Variables have the following columns:
   - "name" is a name of Simple English Wikipedia article,
   - "text" is a full text of the article "name",
   - "out_links" is a list of article names where the article "name" links to.
3. In general you might want to store the counted results in a file before you do the similarity computations and all the research for the third and fourth subtask. Doing all this counting and preperation might allready take quite some runtime.
4. When computing the sparse tf-idf vectors you might allready want to store the eukleadan length of the vectors. otherwise you might discover runtime issues when computing the length again for each similarity computation.
5. Finding the top similar articles for a given article id requires you to compute the similarity of the given article with comparison to all the other known articles and extract the top 5 similarities. Bare in mind that these are quite a lot of similarity computations! You can expect a runtime to find the top similar articles with respect to one of the methods to be up to 10 seconds. If it takes significant longer then you probably have not used the best data structures handle your data.
6. **Even though many third party libraries exist to do this task with even less computational effort those libraries must not be used.**
7. You can find more information about basic usage of pandas DataFrame in pandas documentation.
8. Here are some usefull examples of operations with DataFrame:

Important Notes

Submission

- Solutions have to be checked into the github repository. Use the directory name `groupname/assignment8/` in your group's repository.
- The name of the group and the names of all participating students must be listed on each submission.
- Solution format: all solutions as *one* PDF document. Programming code has to be submitted as Python code to the github repository. Upload *all* `.py` files of your program! Use `UTF-8` as the file encoding. *Other encodings will not be taken into account!*
- Check that your code compiles without errors.
- Make sure your code is formatted to be easy to read.
  - Make sure you code has consistent indentation.
  - Make sure you comment and document your code adequately in English.
  - Choose consistent and intuitive names for your identifiers.
- Do *not* use any accents, spaces or special characters in your filenames.

Acknowledgment

This latex template was created by Lukas Schmelzeisen for the tutorials of "Web Information Retrieval".

LaTeX

Currently the code can only be build using LuaLaTeX, so make sure you have that installed. If on Overleaf, there's an error, go to settings and change the LaTeXengine to `LuaLaTeX`.