

Assignment 3

Centroid Based Document Summarization

Task) Derive a 10-sentence document summary from the 850-sentence input document, 'docs4sum.txt'.

Approach)

- I) Analyze input data
- II) Apply text pre-processing
- III) Cluster document into 10 groups
- IV) Extract most significant sentence from each cluster
- V) Return 10 sentences as the document summary

The above procedure will be implemented in Python. NLTK will be used for text pre-processing in Part II, and Sci-Kit Learn will be used for clustering in Part III.

Part I) Analyze input data

Input data 'docs4sum.txt' contains 850 sentences, each of which separated by new line. There are several sentences that provide no significant meaning, some of which include only punctuation and others that appear to be date stamps.

Part II) Apply text pre-processing

The input data must be prepared for clustering. As noted in Part I, the data set contains several sentences that provide little significant meaning, and as anomalies will negatively impact clustering. Once the input document is read and split into an array of sentences, we remove all documents that have a length less than 35 characters. Removing all sentences less than 35 characters reduced the original 850 sentence document down to 786 sentences. The number 35 was selected through analysis of input data. It successfully removes sentences that contain punctuation, date stamps, and that are just too short to contribute significant meaning to the summary.

Next the sentences are tokenized, stripped of stop words, and all tokens are stemmed using Porter Stemmer algorithm. NLTK's stop word list and Porter Stemmer algorithm are used to complete these steps. The input data is now ready for clustering.

```
43
44 #####
45 # PRE-PROCESSING
46 #####
47
48 # split document into sentences and strip whitespace (delimited by line)
49 sentences = raw_data.split('\n')
50 sentences = map(lambda sentence: sentence.strip(), sentences)
51
52 # remove sentences that do not contribute meaning by assuming short sentences have less meaning
53 sentences = removeShortDocs(sentences, min_sentence_length)
54
55 # remove stop words from all sentences
56 processedSentences = removeStopWords(sentences, nltk_stop_words)
57
58 # stem all tokens of all sentences
59 processedSentences = stemSentences(sentences, ps)
60
61
```

Part III) Cluster document into 10 groups

The K Means clustering algorithm will be used to separate the input data into 10 clusters. Alternative hierarchical clustering methods are helpful when we do not know the desired cluster count K, but since cluster count K is known in advance to be 10, K Means clustering will be used in this example.

A TF-IDF matrix is calculated from the pre-processed sentences returned by Part II. This TF-IDF matrix is then used to fit a K Means Cluster algorithm. The clustering algorithm returns an array of integer labels in the range [0, 9], denoting the cluster assignment for each original sentence.

```
61
62 #####
63 # Apply K Means Clustering
64 #####
65
66 # create tfidf matrix from the processed sentences
67 vectorizer = TfidfVectorizer()
68 tfidf_matrix = vectorizer.fit_transform(processedSentences)
69
70 # cluster our tokenized sentences into 10 groups
71 kMeansCluster = KMeans(n_clusters=cluster_count)
72 kMeansCluster.fit(tfidf_matrix)
73 clusters = kMeansCluster.labels_.tolist()
74
75
```

Before proceeding, the K Means Cluster results are organized to simplify future tasks. Two dictionaries are created to organize the clustering results.

1. A sentence dictionary that stores the original sentence, the stemmed sentence, and the cluster that the sentence was assigned to.
2. A cluster dictionary that contains 10 entries. Each entry points to an array of sentences that have been assigned to given cluster.

```

75
76 #####
77 # Organize Cluster Results
78 #####
79
80 # Create new dictionary that tracks which cluster each sentence belongs to
81 # keeps copy of original sentences and stemmed sentences
82 # sentenceDictionary { idx: { text: String, stemmed: String, cluster: Number } }
83 sentenceDictionary = {}
84 ▼ for idx, sentence in enumerate(sentences):
85     sentenceDictionary[idx] = {}
86     sentenceDictionary[idx]['text'] = sentence
87     sentenceDictionary[idx]['cluster'] = clusters[idx]
88     sentenceDictionary[idx]['stemmed'] = processedSentences[idx]
89
90 # Create new dictionary that contains 1 entry for each cluster
91 # each key in dictionary will point to array of sentences, all of which belong to that cluster
92 # we attach the index to the sentenceDictionary object so we can recall the original sentence
93 clusterDictionary = {}
94 ▼ for key, sentence in sentenceDictionary.items():
95     if sentence['cluster'] not in clusterDictionary:
96         clusterDictionary[sentence['cluster']] = []
97     clusterDictionary[sentence['cluster']].append(sentence['stemmed'])
98     sentence['idx'] = len(clusterDictionary[sentence['cluster']]) - 1
99
100

```

Some tracking is required in order to maintain a reference to the original sentence. For this reason, the sentences position in the cluster sentence array is stored in the sentence dictionary for future reference (line 98).

Part IV) Extract most significant sentence from each cluster

To find the most meaningful sentence in each cluster, cosine similarity scores will be calculated for every cluster's sentence-sentence pair. The resulting cosine similarity matrix is used to identify the single sentence that has the highest similarity with all other documents. The scores of each row in cosine similarity matrix are summed, providing a similarity score for each sentence. The sentence with the highest overall score is then stored, to be used in Part V.

Cosine Similarity was used to determine sentence similarity because it is independent of document size. Other similarity measures, like Euclidean distance, will be negatively affected by the varying sentence sizes.

```
100
101 #####
102 # Calculate Cosine Similarity Scores
103 #####
104
105 # For each cluster of sentences,
106 # Find the sentence with highest cosine similarity over all sentences in cluster
107 maxCosineScores = {}
108 ▼ for key, clusterSentences in clusterDictionary.items():
109     maxCosineScores[key] = {}
110     maxCosineScores[key]['score'] = 0
111     tfidf_matrix = vectorizer.fit_transform(clusterSentences)
112     cos_sim_matrix = cosine_similarity(tfidf_matrix)
113 ▼ for idx, row in enumerate(cos_sim_matrix):
114     sum = 0
115     for col in row:
116         sum += col
117 ▼ if sum > maxCosineScores[key]['score']:
118     maxCosineScores[key]['score'] = sum
119     maxCosineScores[key]['idx'] = idx
120
121
```

Part V) Return 10 sentences as the document summary

In Part IV a sentence of high similarity was extracted from each cluster and stored. Finally, these 10 sentences are combined to construct the desired document summary.

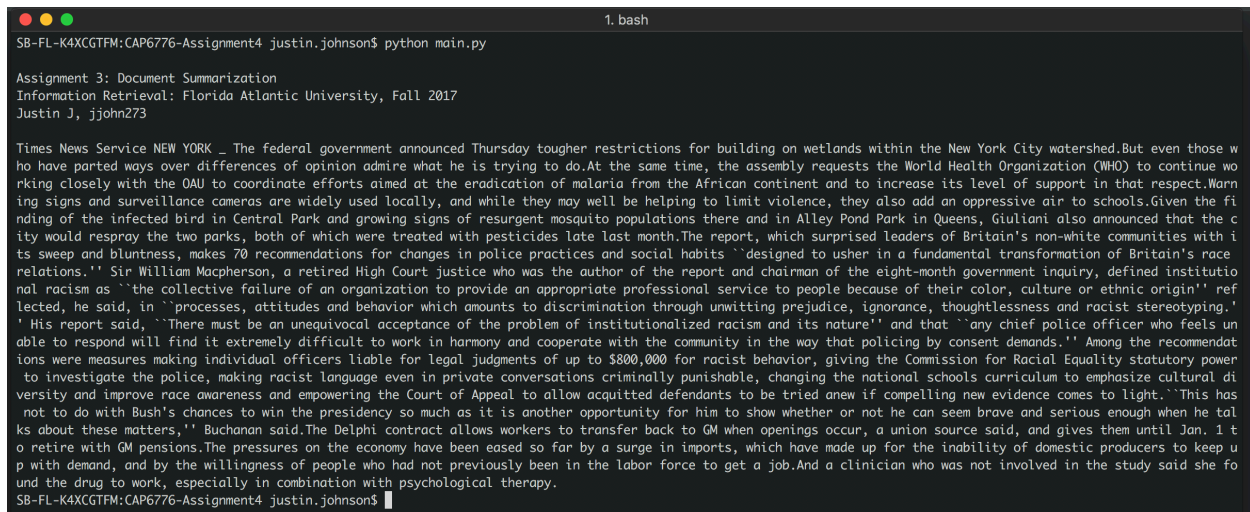
For each of the 10 resulting sentences, their position index is stored in a results array. This array is then sorted in ascending order so that the document summary sentences will be listed in the same order as the original document. This technique was applied to maintain order with the original document.

Finally, the document summary is returned as a concatenation of the 10 sentences extracted in Part IV.

```
121
122
123 #####
124 # Construct Document Summary
125 #####
126
127 # for every cluster's max cosine score,
128 # find the corresponding original sentence
129 resultIndices = []
130 i = 0
131 ▼ for key, value in maxCosineScores.items():
132     cluster = key
133     idx = value['idx']
134     stemmedSentence = clusterDictionary[cluster][idx]
135     # key corresponds to the sentences index of the original document
136     # we will use this key to sort our results in order of original document
137 ▼ for key, value in sentenceDictionary.items():
138 ▼     if value['cluster'] == cluster and value['idx'] == idx:
139         resultIndices.append(key)
140
141 resultIndices.sort()
142
143 # Iterate over sentences and construct summary output
144 result = ''
145 ▼ for idx in resultIndices:
146     print idx
147     result += sentences[idx]
148
149
150 print result
151
152
```

Results)

Screenshot of Console Output



```
SB-FL-K4XCCTFM:CAP6776-Assignment4 justin.johnson$ python main.py

Assignment 3: Document Summarization
Information Retrieval: Florida Atlantic University, Fall 2017
Justin J, jjohn273

Times News Service NEW YORK _ The federal government announced Thursday tougher restrictions for building on wetlands within the New York City watershed. But even those who have parted ways over differences of opinion admire what he is trying to do. At the same time, the assembly requests the World Health Organization (WHO) to continue working closely with the OAU to coordinate efforts aimed at the eradication of malaria from the African continent and to increase its level of support in that respect. Warning signs and surveillance cameras are widely used locally, and while they may well be helping to limit violence, they also add an oppressive air to schools. Given the finding of the infected bird in Central Park and growing signs of resurgent mosquito populations there and in Alley Pond Park in Queens, Giuliani also announced that the city would respray the two parks, both of which were treated with pesticides late last month. The report, which surprised leaders of Britain's non-white communities with its sweep and bluntness, makes 70 recommendations for changes in police practices and social habits "designed to usher in a fundamental transformation of Britain's race relations." Sir William Macpherson, a retired High Court justice who was the author of the report and chairman of the eight-month government inquiry, defined institutional racism as "the collective failure of an organization to provide an appropriate professional service to people because of their color, culture or ethnic origin" reflected, he said, in "processes, attitudes and behavior which amounts to discrimination through unwitting prejudice, ignorance, thoughtlessness and racist stereotyping." His report said, "There must be an unequivocal acceptance of the problem of institutionalized racism and its nature" and that "any chief police officer who feels unable to respond will find it extremely difficult to work in harmony and cooperate with the community in the way that policing by consent demands." Among the recommendations were measures making individual officers liable for legal judgments of up to $800,000 for racist behavior, giving the Commission for Racial Equality statutory power to investigate the police, making racist language even in private conversations criminally punishable, changing the national schools curriculum to emphasize cultural diversity and improve race awareness and empowering the Court of Appeal to allow acquitted defendants to be tried anew if compelling new evidence comes to light. "This has not to do with Bush's chances to win the presidency so much as it is another opportunity for him to show whether or not he can seem brave and serious enough when he talks about these matters," Buchanan said. The Delphi contract allows workers to transfer back to GM when openings occur, a union source said, and gives them until Jan. 1 to retire with GM pensions. The pressures on the economy have been eased so far by a surge in imports, which have made up for the inability of domestic producers to keep up with demand, and by the willingness of people who had not previously been in the labor force to get a job. And a clinician who was not involved in the study said she found the drug to work, especially in combination with psychological therapy.
```

The above screenshot displays the output of the algorithm described throughout this project. Results will vary each iteration of the algorithm, because K Means clustering results are partially dependent on the initialization seeds. In addition, the K Means algorithm may get stuck at local minimum when converging, leading to varying results. The document summary in the above screenshot can be found on the following page.

Document Summary Text

Times News Service NEW YORK _ The federal government announced Thursday tougher restrictions for building on wetlands within the New York City watershed. But even those who have parted ways over differences of opinion admire what he is trying to do. At the same time, the assembly requests the World Health Organization (WHO) to continue working closely with the OAU to coordinate efforts aimed at the eradication of malaria from the African continent and to increase its level of support in that respect. Warning signs and surveillance cameras are widely used locally, and while they may well be helping to limit violence, they also add an oppressive air to schools. Given the finding of the infected bird in Central Park and growing signs of resurgent mosquito populations there and in Alley Pond Park in Queens, Giuliani also announced that the city would respray the two parks, both of which were treated with pesticides late last month. The report, which surprised leaders of Britain's non-white communities with its sweep and bluntness, makes 70 recommendations for changes in police practices and social habits ``designed to usher in a fundamental transformation of Britain's race relations." Sir William Macpherson, a retired High Court justice who was the author of the report and chairman of the eight-month government inquiry, defined institutional racism as ``the collective failure of an organization to provide an appropriate professional service to people because of their color, culture or ethnic origin" reflected, he said, in ``processes, attitudes and behavior which amounts to discrimination through unwitting prejudice, ignorance, thoughtlessness and racist stereotyping." His report said, ``There must be an unequivocal acceptance of the problem of institutionalized racism and its nature" and that ``any chief police officer who feels unable to respond will find it extremely difficult to work in harmony and cooperate with the community in the way that policing by consent demands." Among the recommendations were measures making individual officers liable for legal judgments of up to \$800,000 for racist behavior, giving the Commission for Racial Equality statutory power to investigate the police, making racist language even in private conversations criminally punishable, changing the national schools curriculum to emphasize cultural diversity and improve race awareness and empowering the Court of Appeal to allow acquitted defendants to be tried anew if compelling new evidence comes to light. ``This has not to do with Bush's chances to win the presidency so much as it is another opportunity for him to show whether or not he can seem brave and serious enough when he talks about these matters," Buchanan said. The Delphi contract allows workers to transfer back to GM when openings occur, a union source said, and gives them until Jan. 1 to retire with GM pensions. The pressures on the economy have been eased so far by a surge in imports, which have made up for the inability of domestic producers to keep up with demand, and by the willingness of people who had not previously been in the labor force to get a job. And a clinician who was not involved in the study said she found the drug to work, especially in combination with psychological therapy.