Assignment 1 - Beyond word count

Anirudh Narasimhamurthy(u0941400) and Soumya Smruti Mishra()

September 7, 2015

1 Word Transition counts

In this part of the assignment we extend the basic word count example to calculate count of word pairs or 2-grams or bigrams. The data structure used to compute the number of times word2 follows word1 in the given file was implemented using **tuples** in Python.

The input file had a lot of punctuations and leading and trailing spaces. If we consider the input file as such and pass it to the sc.textFile() method, the count of the most commonly occurring word pairs has variations of the same word pair with differences in punctuation. A brief summary of our approach to find the most common word pairs in the document is provided below:

- Use the textFile() method to create a RDD of input file as lines.
- Map the lines to individual words by using an appropriate lambda function inside map()
- Create tuples of word pairs by applying flatMap() over the mapped wordsRDD
- Filter the tuples or word-pairs with word length less than 5 characters using the filter function
- Apply map() and reduceByKey() over the filteredRDD to obtain the appropriate word pair counts.
- Apply the takeOrdered() function on the resulting RDD from the previous step and provide the appropriate parameter values to obtain the ten most frequently occurring word pairs.

The results after running our program on the input file are tabulated in Table 1:

Clearly the results in the top 10 contain a variation of the same word pair in 3 different places owing to the punctuation. So we also tweaked our code to remove all the punctuation, leading and trailing spaces and convert the entire text to lowercase. The remove punctuation was done immediately after loading the document using sc.textFile and so the filtered/cleaned up text document was then passed to the rest of the functions.

On running the input doc through the punctuation remover, the results of the 10 frequently occurring word pairs are tabulated in Table 2 :

Ten most frequently occurring word pairs (len(word) > 4)

Word Pair	Count
(u'Prince', u'Andrew')	631
(u'United', u'States')	229
(u'Prince', u'Andrew,')	163
(u'Prince', u'Vasili')	140
(u'Prince', u'Andrew.')	97
(u'Project', u'Gutenberg-tm')	86
(u'Prince', u"Andrew's")	76
(u'United', u'States,')	68
(u'takes', u'place')	67
(u'Project', u'Gutenberg')	66

Table 1: Top 10 most frequently occurring word pairs for input document

Word Pair	Count
(u'prince', u'andrew')	907
(u'united', u'states')	392
(u'prince', u'vasili')	178
(u'project', u'gutenbergtm')	104
(u'project', u'gutenberg')	99
(u'sherlock', u'holmes')	99
(u'mademoiselle', u'bourienne'):	91
(u'takes', u'place')	90
(u'prince', u'andrews'):	79
(u'marya', u'dmitrievna')	76

Table 2: Top 10 most frequently occurring word pairs for after removing punctuation and trailing spaces in input document

2 Text similarity

Part A : Jaccard Distance/Similarity

For computing the Jaccard distance across all pairs of documents, we first need to find all the words in one document and then do a cartesian with words in the other document.

The similarity matrix in this case would be a 451 x 451 matrix, assuming we are taking all the 451 files in the corpus.

The diagnol entries of the matrix would be 1 and the matrix would be a symmetric matrix since similarity between ai and ak is same as ak and ai.

We were able to run our code on all the 451 documents and obtain the Jaccard Similarity matrix. But plotting using matplotlib was taking or running for a really long time. Since matplotlib expects the input to be in a numpy array, we had to filter and transfer the results from the RDD into individual numpy arrays.

A plot of Jaccard similarity between the different documents is shown below :

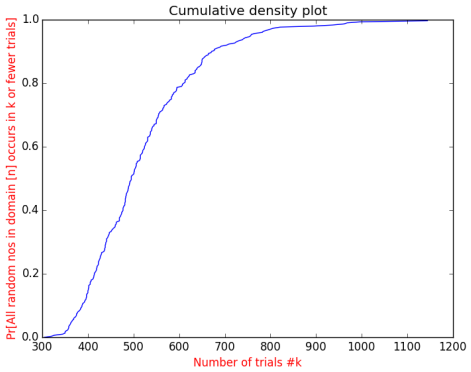
Part B : Cosine Distance/Similarity

To compute the Cosine Distance, we first found the 1000 most common words across the set of all documents based on the word count.

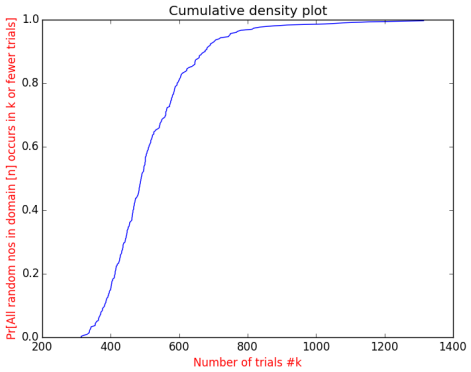
The most common 1000 words were used as the basis for building up the individual feature vectors where individual entries in the the feature vector represents the number of times the word in the common list appears in the individual document.

- If the word in the common list did not occur in the individual documents, then its corresponding value in the vector was set as 0.
- We again used the **cartesian** function to obtain all pair of documents similarity and processed and stored the results in a similar fashion to the Jaccard similarity results.

A plot of Jaccard similarity between the different documents is shown below :



(a) Cumulative density plot for Coupon Collector simulation



(b) Cumulative density plot for Coupon Collector simulation

Text file containing the similarity matrices has also been added to the zipped project folder submitted with this assignment.