CHD Lab
Assignment 1

Question

A bag-of-words approach is a simplifying representation used in NLP where text is represented in terms of words and their frequency.

Perform a bag-of-words analysis on the given essays.

Example Method

1. Load the raw text
2. Split into tokens
3. Convert all tokens to lower case
4. Remove punctuation from each token
5. Remove tokens that are not alphabetic
6. Remove stop words
7. Stem all words
8. Get a count of all words

Code Snippets

- To load an entire file into memory :

```
filename = "essay.txt"
infile = open(filename, "r")
text = infile.read()
infile.close()
```

- To tokenize text into words:

```
from nltk.tokenize import word_tokenize
tokens = word_tokenize(text)
```

- To remove punctuation from a string:

```
import re
token = re.sub(r'[^\w\s]', '', token)
```

- To remove tokens that are not alphabetic from a list of strings:

```
words = [word for word in tokens if word.isalpha()]
```

- To convert a word to lower case:

```
word = word.lower()
```

- To get list of stopwords (ex : the/a/is)

```
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
```

- To stem a word:

```
from nltk.stem import SnowballStemmer
snowball = SnowballStemmer('english')
stemmed_word = snowball.stem(word)
```

Helpful Tutorials

- How to Clean Text for Machine Learning with Python
- Count frequencies of all elements in array in Python

Submission Format

Submit a zip file rollnumber.zip with your python code(rollnumber.py) and three files of the form essayname.txt, each containing frequency counts in the format :

```
word 1 : count 1
word 2 : count 2
..
..
```