Homework 1: Topic Modeling

Start Assignment

Due Jan 24 by 2:30pm   **Points** 5   **Submitting** a file upload

Goal

The goal of this homework is to introduce Topic Modeling, and try to come up with some ideas for a way to determine the best way to describe the contents of a text document.

Setup and Requirements

- Work individually on this homework.
- This program can be written in whatever language that the TAs are comfortable with. (For anything other than Python, Java, and C++ please check with the TAs before you submit)

Your Task

Write a program named `topicmodel.py` (or equivalent in whatever language you're using) that does the following:

Input: a text file (the filename is passed in as the first and only command-line parameter)

Output: the *five* words (printed out, one per line) that best summarize the contents of the text file

For example, here is a trivial and terrible first idea:

```python
import sys

def best_words(f, maxLength):
    words = []
    for line in f:
        for word in line.split():
            if len(words) < maxLength:
                words.append(word)
    return words

def main():
    f = open(sys.argv[1])
    for word in best_words(f, 5):
        print(word)
    f.close()

if __name__ == '__main__':
    main()
```

The output that I get when I run it on Franklin Delano Roosevelt's first inaugural address found here ⬀ is as follows:

```
----
% python3 topicmodel.py fdr-firstinaugural.txt
I
am
certain
that
my
----
```

The main point of this assignment is to come up with a non-silly algorithm to solve the problem.

Spend a little time thinking about what decisions you are making, and whether they make sense.

Don't spend more than 3 hours on this problem. Come up with something that is at least a little better than my program.

Come to class on Monday prepared to talk about your approach.

Feel free to use my program as a starting point.

◄ Previous   Next ►