# COMP 370 Homework 11 – Using TF-IDF

Assigned Nov 15, 2024  
Due Nov 22, 2024 @ 11:59 PM

In this assignment, we’re going to use TF-IDF to compare the language used by 6 different subreddits.

## Task 1: Collect your reddit data

Select 6 subreddits – any 6 you’d like, presumably on topics that you like or are interested in. For example, I might select r/mcgill, r/montreal, r/programming, r/marvel, r/mylittlepony, r/3Dprinting. Note that, generally speaking, the more related the subreddits are, the more helpful TF-IDF will be.

Once you’ve selected them, download the latest 200 posts from each. Put them in files: <subreddit\_name>.json

## Task 2: Compute naïve most-frequent-word list

Write a script naive\_frequent\_word\_list.py that computes the top 10 words that appear in the 200 post titles based on absolute frequency. It should count the number of times each word appears in the titles and then rank order them. The script should run as follows:

python build\_naive\_frequent\_word\_list.py -o <word\_counts.json> <the list of input files>

Note that your script should accept an arbitrary number of word lists. Bear in mind that you need to clean the titles since they’ll contain punctuation and such… which aren’t actually a part of a word.

The output file should be a dictionary with the following form:

{

“<input\_file\_1>”: [

[“<word1>”, <# of times the word1 is used>],

[“<word2>”: <# of times the word2>,

…

],

“input\_file\_2”: [

…

]

...

}

## Task 3: Add stop word removal to your naïve script

Add stop word removal to your script such that it can be run as follows:

python build\_naive\_frequent\_word\_list.py -o <word\_counts.json> [-s <stop\_word\_file>] <the list of input files>

The stop word file is optional. If it’s specified, then all the words in the stop word file list are removed from the candidate list of top words.

Find and download a good stopword list – there are many on the internet.

## Task 4: Implement TF-IDF script

Implement a separate script that computes the top 10 words for each input file based on their TF-IDF score (where each subreddit is considered the “document”). It should run as follows:

python build\_tfidf\_word\_list.py -o <word\_scores.json> [-s <stop\_word\_file>] <the list of input files>

The contract is the same except that your output file contains the tf-idf scores rather than the # of times they’re used.

## Task 5: Compare top word list methods

Use your scripts to compute the following top word lists:

* naive\_no\_stops.json: your naïve method with no stop words
* naive\_with\_stops.json: your naïve method with stop words
* tfidf\_no\_stops.json: your tf-idf method with no stop words
* tfidf\_with\_stops.json: your tf-idf method with stop words

Compare the top word lists (you might need to write some additional scripts to make that easier to do). Answer the following questions.

1. What is the impact of including a stop word list?
2. What differences do you observe with TF-IDF?
3. Which method produces the best list?

## Submission Instructions

Submit a zip file hw11.zip containing the following:

* All 6 files from Task 1 (named as indicated) – they’ll all be .json, but besides that everyone will have different ones.
* build\_naive\_frequent\_word\_list.py
* build\_tfidf\_word\_list.py
* All 4 files from Task 5 (named as indicated)
* analysis.md – a file containing your answers to the questions posed in Task 5.

Note – I’m not asking for your stop words file since they’ll all be different and possibly different formats.