**Data**

The file (tweets.txt) contains 100,000 tweets about the movie “ **American Sniper** ”. These are unprocessed real tweets.

**The Problem**

As you may know the movie was directed by **Clint Eastwood** , starring **Bradley Cooper** and was based on the life story of **Chris Kyle** .

Your **first task** is to break the twitter hashtags into proper words, for example:

#goldenglobes -> golden globes,

#siennamiller -> sienna miller ...Example 1

Your **second task** is to find out what people have said about these three personalities ( **Clint Eastwood, Bradley Cooper** and **Chris Kyle** ), in other words find out the top trending topics about these people in the current dataset.

For example, consider these 3 tweets about **Sienna Miller** from the same dataset:

● Sienna Miller is stunning!!! #GoldenGlobes She brought the heat in @AmericanSniper!!!

● @AmericanSniper and @FoxcatcherUK actress #SiennaMiller looks stunning in #Forevermark diamond jewelry. http://t.co/2Y8JvJ8KlF

● Watching #AmericanSniper awesome acting by @BradleyCooperUS and @SiennaMillerr playing the life of Chris Kyle and Taya a ten

So for Sienna Miller one trending topic can be “looks stunning” or just “stunning” and another can be “awesome acting”, this can be represented as follows:

Sienna Miller

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ ... **Example 2**

stunning (2 tweets)

awesome acting (1 tweet)

Remember these are not the only trends which can be extracted from the above 3 tweets. Other trends may include Golden Globes, diamond jewelry, Forevermark, FoxcatcherUK etc. Use your own judgement to figure out meaningful trends.

**Instructions and Expected Results**

● **Avoid** using any 3rd party library for the main logic, but you can use libraries for the trivial tasks like

stemming, tagging, tokenization etc. You can also use a 3rd party english dictionary, if you need one.

● You are free to choose any mainstream language.

● Do not use Regex excessively, we are looking for an NLP approach for the first problem.

● Words with similar meaning should be considered same, i.e., stunning, superb, wonderful and awesome should be treated as one word, you can pick which ever word you like.

looked awesome (23 tweets)

looked wonderful (7 tweets)

should be clubbed as looked wonderful (30 tweets) or looked awesome (30 tweets)

● Final top trends should be relevant and meaningful.

● Show 5 top trends per person. The tweet count should be in decreasing order. (use Example 2 format)

● A trend can be a single word ( *stunning* ) or may comprise of multiple words ( *looks stunning* ) as shown in Example 2.

● Upload your code to github and share the url or you can zip and email it back. We also expect a small writeup explaining your approach and code, further improvements and details about 3rd party libs you have used.