* What is my project now since OLDA is in Spark?
* What do I do for the midterm report?
  + Define the problem im interested in, identifying trending topics or something in Twitter
  + Describe background – in dataset how much data and whatnot
  + Describe olda: describe lda, how to do lda in online manner
  + Run olda on some of data, present preliminary results
  + Write short section on things we can do for rest of the project, how to do olda to trending topics
* What is the input for olda spark example / like how do we get that from text?
* How do we get Z values?

Run LDA w/ 10, 20, 50, 100

Figure out what we are learning for Alpha, check for each LDA,

For 10 most often used topics look at most often appearing words

Take vectors theta, add them together

Find the topics with the highest alpha, compare them to thetas

Check topics

Write for a classmates understanding

Look up period of time im looking at and check the media stories to see what I am expecting to see

(156.45gb during, 22.11gb after) 134gb in middle term files to run lda on the whole dataset

Took an hour to get there (timed using Avengers Infinity War timestamps since it broke)

OSError: [Errno 28] No space left on device

"""

PicklingError: Could not pickle the task to send it to the workers.

Blei et al. suggests that one should remove common stop words in a document before running Latent Dirichlet Allocation [2]. Stop words are words that are very common and do not alter the meaning of a sentence [6], typical stop words include ”the”, ”in” and ”so” etc. Kim et al. automatically filtered out words from the vocabulary that were present in more than 50% or less that 5% of the documents [6]. They state that it was an effective way of removing both stop words, misspelled words and non-words. In this thesis a static list of common stop-words2 was used as suggested by Blei et al.