Homework 1. Text Mining for Social Sciences

Ivan Vallejo Vall, Daniel Velasquez Vergara, Saurav Poudel

April 20, 2017

1 HOMEWORK 2

1.1 Text Mining for Social Sciences

1.2 Ivan Vallejo Vall, Daniel Velasquez Vergara, Saurav Poudel, Viviana Rosales

1.2.1 11 May 2017

1.2.2 Exercise 1

We start by reusing the code developed in homework 1 to create the document term matrix of the State of the Union addresses.

This time, however, we take as a document the whole speech for a given year, instead of each paragraph as we did in homework 1. LDA allows for multiple topic allocation per document. Each paragraph will probably have a single topic and therefore we would not take much advantage of multiple topic allocation, whereas at the aggregate level (year) we will certainly have multiple topics.

Moreover, it is also more relevant for the analysis to have an aggregate measure at the year level of topic evolution, rather than a detailed analysis per paragraph.

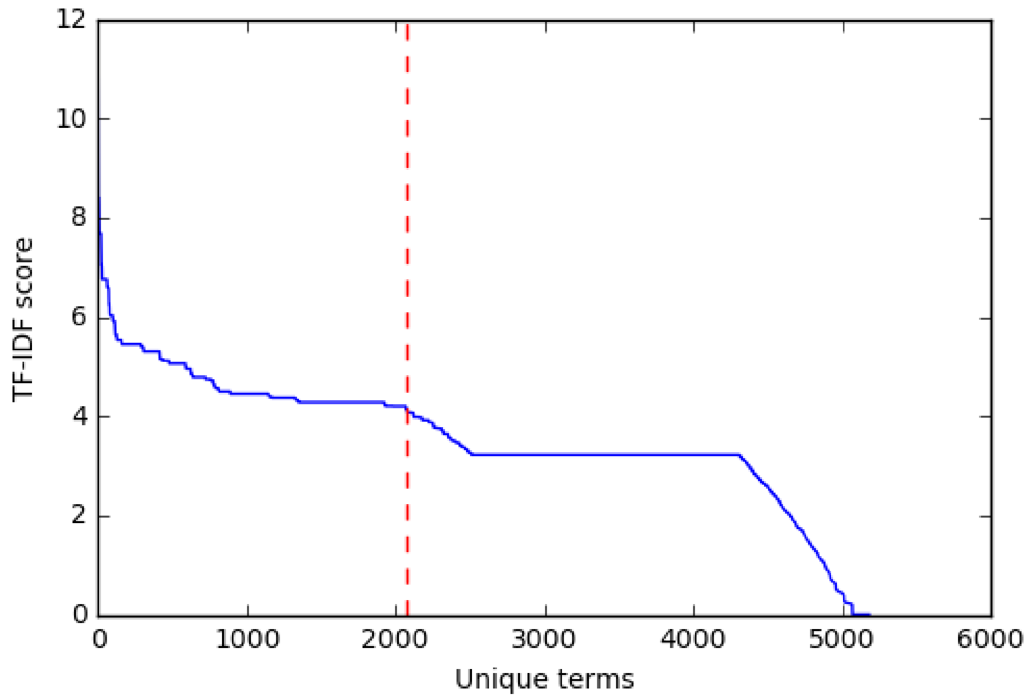
The following code creates the desired document term matrix, taking speeches starting from the year 1990 and applying a TF-IDF cut-off as specified in the figure below. For longer time periods and/or more terms selected, the procedure of the following steps would be the same, it would just require extra processing time.

Note: the file 'speech_data_extend.txt' is needed to run the code, plus the nl corpora, which can be downloaded by typing nl.download().

```
In [2]: #import packages
        import numpy as np
        import matplotlib.pyplot as plt
        import nltk as nl
        from nltk.tokenize import word_tokenize
        import pandas as pd
        from stop_words import get_stop_words
        from nltk.stem.porter import PorterStemmer
        import operator
        # Download corpora if necessary: nl.download()

        # Start analysis from this year
        year = 1990
        span = 2014-year+1

        # Import state-of-the-union speech
        text_raw = pd.read_csv('./speech_data_extend.txt', sep='\t')
```

```python
# Consider addresses from 1970
text_data = text_raw.loc[text_raw['year']>=year, :]

# Reconstitute full speech for each year
text_year = pd.DataFrame(index=range(span), columns=['speech','year'])
for i in range(span):
    text_year['speech'][i]=' '.join(text_data['speech'][text_data['year']==year+i])
    text_year['year'][i]= year+i

#Processing of the data
stop_words = get_stop_words('en')
st = PorterStemmer()
docs = pd.Series(np.zeros(text_year.shape[0]))
tokens = [] #List of all words.

for i, line in enumerate(text_year['speech']):
    #Tokenize the data:
    doc_i = word_tokenize(line.lower())
    #Remove non-alphabetic characters:
    doc_i = [tok for tok in doc_i if tok.isalpha()]
    #Remove stopwords using a list of your choice:
    doc_i = [tok for tok in doc_i if tok not in stop_words]
    #Stem the data using the Porter stemmer:
    doc_i = [st.stem(tok) for tok in doc_i]

    tokens.extend(doc_i)
    docs.iloc[i] = doc_i

# Corpus-level tf-idf score for every term, and choose a cutoff below which to remove
words.
unique_words = np.unique(tokens)
lw = len(unique_words) # Number of words
ld = len(docs) # Number of documents

word_count = nl.FreqDist(tokens)
tf = {k: 1+np.log(v) for k, v in word_count.items()}
df = {k: np.sum(list(map(lambda x: k in x, docs))) for k in word_count.keys()}
idf = {k: np.log(ld/v) for k, v in df.items()}
tfidf = {k : v * tf[k] for k, v in idf.items() if k in tf}

# Based on the ranking we select 500 words with highest tf-idf
# 1st we get the rank
rank = sorted(tfidf.items(), key=operator.itemgetter(1), reverse=True)
cutoff = rank[2000][1] -0.0001
# 2nd apply the cut-off
selected_words = {k: v for k, v in tfidf.items() if v>cutoff}
ls = len(selected_words) # number of selected words

%matplotlib inline
plt.plot([x[1] for x in rank])
plt.axvline(ls, color='red',linestyle='dashed')
plt.xlabel("Unique terms")
plt.ylabel("TF-IDF score")

print("\n Number of unique words: %d" %lw)

print("\n Number of selected words (cutoff %3.1f tf-idf): %d" %(cutoff,ls))

#Document-term matrix using words selected using the tf-idf score.
X = pd.DataFrame(np.zeros(shape = (ld, ls)), columns = selected_words.keys())

for w in selected_words.keys():
    X[w] = list(map(lambda x: x.count(w), docs))
```

```
Number of unique words: 5183

Number of selected words (cutoff 4.2 tf-idf): 2070
```



We initialize the Gibbs sampler by setting the parameters (α, η, #iterations and #topics) and initializing the matrices we need to run it:

- $\boldsymbol{\theta}_d$: document specific mixing probabilities, D x K matrix.
- $\boldsymbol{\beta}_k$: topic specific term probabilities, K x V matrix.
- $\boldsymbol{z}_{d,n}$: topic allocation to each term of each document, D x V matrix.
- $\boldsymbol{n}_{d,k}$: number of words in document d that have topic allocation k, D x K matrix.
- $\boldsymbol{m}_{k,v}$: number of times topic k allocation variables generate term v, K x V matrix.

where

- **D:** number of documents. Dependent on the starting year and the aggregation. In our case, we cover the period 1990-2014 (25 years) and the level of aggregation is the whole speech of a given year, therefore D = 25.
- **K:** number of topics. A parameter of the Gibbs sampler. We try with 5 topics to facilitate the interpretation of the results (the more number of topics, the more difficult to associate each one with a given external phenomena).
- **V:** number of terms. Dependent on the previous step TF-IDF cut-off applied. In our case, V = 2'070.

As proposed by Griffiths and Steyvers, we set $\eta = 200/V \approx 0.1$ and $\alpha = 50/K = 10$. In order to ensure that the algorithm has enough time to converge, we set #iterations = 12'000.

In [5]:
```python
#import packages
import numpy as np
import pandas as pd
from random import randint
import collections

# parameters document term matrix
D = X.shape[0]
V = X.shape[1]

# parameters gibbs sampler
topics = 5
alpha = 10
eta = 0.1
iterations = 12000

# initialize randomly (i.e. equal prob) matrix theta d
# topics numbered from 0 to k-1
theta_docs = 1/topics * np.ones(shape = (D, topics))

# initialize randomly (i.e. equal prob) matrix beta k
# topics numbered from 0 to k-1
beta_terms = 1/V * np.ones(shape = (topics,V))

# initialize the matrix z d,n
# topics numbered from 1 to k (cannot use 0 because it is used for non occurrences)
TA_terms = X.as_matrix()
for doc in range(D):
        for term in TA_terms[doc,:].nonzero()[0]:
            TA_terms[doc,term] =
1+np.random.multinomial(1,theta_docs[doc,:],size=1).argmax()

# initialize matrix n d,k
# topics numbered from 0 to k-1
TA_doc = np.zeros(shape = (D, topics))
for i in range(D):
    tmp = collections.Counter(TA_terms[i,:])
    for j in range(topics):
        TA_doc[i,j] = tmp[j+1]

# initialize matrix m k,v
# topics numbered from 0 to k-1
TA_v = np.zeros(shape = (topics,V))
for i in range(V):
    tmp = collections.Counter(TA_terms[:,i])
    for j in range(topics):
        TA_v[j,i] = tmp[j+1]
```

Next we run the GIBBS sampler following the steps outlined in the class slides:

- a) Sample from a multinomial distribution N times for the term-topic allocation: $$P(z_{d,n}|w_{d,n} = v, \boldsymbol{B}, \boldsymbol{\theta_d}) = \frac{\theta_d^k \beta_k^v}{\sum_k \theta_d^k \beta_k^v}$$

b) Update $\boldsymbol{z}_{d,n}$, $\boldsymbol{n}_{d,k}$ and $\boldsymbol{m}_{k,v}$ based on the new topic allocations drawn from the multinomial.

- c) Sample from a Dirichlet D times for the document-specific mixing probabilities: $$P(\boldsymbol{\theta}_d|\alpha, \boldsymbol{z}_d) = Dir\ (\alpha + n_{d,1}, \cdots, \alpha + n_{d,K})$$

- d) Sample from a Dirichlet K times for the topic-specific term probabilities: $$P(\boldsymbol{\beta}_k|\eta, \boldsymbol{w}, \boldsymbol{z}) = Dir\ (\eta + m_{k,1}, \cdots, \eta + m_{k,V})$$

In order to test convergence, we use the perplexity score at the end of each iteration:

$$exp\left[-\frac{\sum_{d=1}^{D}\sum_{v=1}^{V}x_{d,v}log\left(\sum_{k=1}^{K}\hat{\theta}_{d,k}\hat{\beta}_{k,v}\right)}{\sum_{d=1}^{D}N_d}\right]$$

where,

$$\hat{\beta}_{k,v}=\frac{m_{k,v}+\eta}{\sum_{v=1}^{V}(m_{k,v}+\eta)} \qquad \hat{\theta}_{d,k}=\frac{n_{d,k}+\alpha}{\sum_{k=1}^{K}(n_{d,k}+\alpha)}$$

In addition, we keep track of the evolution of topic allocation at each iteration for selected documents. We expect that topic allocation in a given document should become stable as the algorithm converges.

Note: the next chunk of code requires quite some time to run (about 45 min for 12'000 iterations). For a faster test, it can be run for, say, 1'000 iterations by just changing in the previous chunk of code the variable 'iterations' to 1000. The results obtained with 12'000 iterations have been saved and so can be retrieved afterwards.

In [6]:
```
# To control time: import timeit; start_time = timeit.default_timer(); elapsed =
timeit.default_timer() - start_time

# GIBBS sampler
perplexity = np.zeros(iterations)
track = np.zeros(shape = (iterations,topics))
track2 = np.zeros(shape = (iterations,topics))
track3 = np.zeros(shape = (iterations,topics))
track4 = np.zeros(shape = (iterations,topics))
track5 = np.zeros(shape = (iterations,topics))
track6 = np.zeros(shape = (iterations,topics))
X_np = X.as_matrix()
Nd = X_np.sum()

for i in range(iterations):

    #start_time = timeit.default_timer()
    if i % 200 == 0:
        print("Iteration %d " %(i))

    # Sample from a multinomial distribution N times for the term-topic allocation
    for doc in range(D):
        for term in TA_terms[doc,:].nonzero()[0]:
            # sample multinomial to get new topic allocation
            old_z = TA_terms[doc,term]-1
            p_z = np.multiply(theta_docs[doc,:],beta_terms[:,term])
            p_z_sum = p_z.sum()
            new_z = np.random.multinomial(1, p_z / p_z_sum).argmax()

            # update matrices depending on topic allocation
            TA_terms[doc,term] = new_z+1 # update topic-term matrix
            TA_doc[doc,old_z] -= 1 # decrease by one previous topic count in n d,k
            TA_doc[doc,new_z] += 1 # increase by one new topic count in n d,k
            TA_v[old_z,term] -= 1 # decrease by one previous topic count in m k,v
            TA_v[new_z,term] += 1 # increase by one new topic count in m k,v

    # Sample from a Dirichlet D times for the document-specific mixing probabilities
    for doc in range(D):
        theta_docs[doc,:] = np.random.dirichlet(alpha=(alpha+TA_doc[doc,:]))

    # Sample from a Dirichlet K times for the topic-specific term probabilities
    for topic in range(topics):
        beta_terms[topic,:] = np.random.dirichlet(alpha=(eta+TA_v[topic,:]))
```

```python
        # calculate perplexity score
        theta_hat = TA_doc+alpha
        theta_hat = theta_hat / theta_hat.sum(axis=1, keepdims=True)
        beta_hat = TA_v+eta
        beta_hat = beta_hat / beta_hat.sum(axis=1, keepdims=True)
        perplexity[i]=0
        for doc in range(D):
            for term in TA_terms[doc,:].nonzero()[0]:
                perplexity[i] += X_np[doc,term] *
    np.log(np.multiply(theta_hat[doc,:],beta_hat[:,term]).sum())
        perplexity[i] = np.exp(-perplexity[i]/Nd)

        # Keep track of evolution of topic allocation 1
        track[i,:] = theta_docs[0,:]
        track2[i,:] = theta_docs[5,:]
        track3[i,:] = theta_docs[10,:]
        track4[i,:] = theta_docs[15,:]
        track5[i,:] = theta_docs[20,:]
        track6[i,:] = theta_docs[24,:]

        #print("-", end="")
        #elapsed = timeit.default_timer() - start_time
        #print("%4.3f" %elapsed)

    print("Iteration %d " %iterations)
    print("Done Gibbs sampler. Initial perplexity: %.1f ; final perplexity: %.1f"
          %(perplexity[0],perplexity[iterations-1]))
```

```
Iteration 0
Iteration 200
Iteration 400
Iteration 600
Iteration 800
Iteration 1000
Iteration 1200
Iteration 1400
Iteration 1600
Iteration 1800
Iteration 2000
Iteration 2200
Iteration 2400
Iteration 2600
Iteration 2800
Iteration 3000
Iteration 3200
Iteration 3400
Iteration 3600
Iteration 3800
Iteration 4000
Iteration 4200
Iteration 4400
Iteration 4600
Iteration 4800
Iteration 5000
Iteration 5200
Iteration 5400
Iteration 5600
Iteration 5800
Iteration 6000
Iteration 6200
Iteration 6400
Iteration 6600
```

```
Iteration 6800
Iteration 7000
Iteration 7200
Iteration 7400
Iteration 7600
Iteration 7800
Iteration 8000
Iteration 8200
Iteration 8400
Iteration 8600
Iteration 8800
Iteration 9000
Iteration 9200
Iteration 9400
Iteration 9600
Iteration 9800
Iteration 10000
Iteration 10200
Iteration 10400
Iteration 10600
Iteration 10800
Iteration 11000
Iteration 11200
Iteration 11400
Iteration 11600
Iteration 11800
Iteration 12000
Done Gibbs sampler. Initial perplexity: 1777.7 ; final perplexity: 1765.4
```

```
In [24]: # save results to csv files so that we do not need to run the 12'000 iterations
         everytime we open the notebook
         np.savetxt("./results/perplexity.csv",perplexity,delimiter=",")
         np.savetxt("./results/track.csv",track,delimiter=",")
         np.savetxt("./results/track2.csv",track2,delimiter=",")
         np.savetxt("./results/track3.csv",track3,delimiter=",")
         np.savetxt("./results/track4.csv",track4,delimiter=",")
         np.savetxt("./results/track5.csv",track5,delimiter=",")
         np.savetxt("./results/track6.csv",track6,delimiter=",")
         np.savetxt("./results/theta_docs.csv",theta_docs,delimiter=",")
         np.savetxt("./results/beta_terms.csv",beta_terms,delimiter=",")
         np.savetxt("./results/TA_doc.csv",TA_doc,delimiter=",")
         np.savetxt("./results/TA_v.csv",TA_v,delimiter=",")
         np.savetxt("./results/TA_terms.csv",TA_terms,delimiter=",")
```

Next, we monitor the evolution of perplexity as well as the results of topic allocation for several documents. The charts below show the results using a moving average to smooth them, as otherwise oscillation make the charts difficult to read.

We note that perplexity does not clearly improve with the number of iterations, although there may be a marginal downward linear trend (see red line).

We also remark that the topic allocation of each speech does not converge with the number of iterations to a clear stable pattern. This also casts some doubt on the robustness of the results we may derive from the topic allocation matrix.

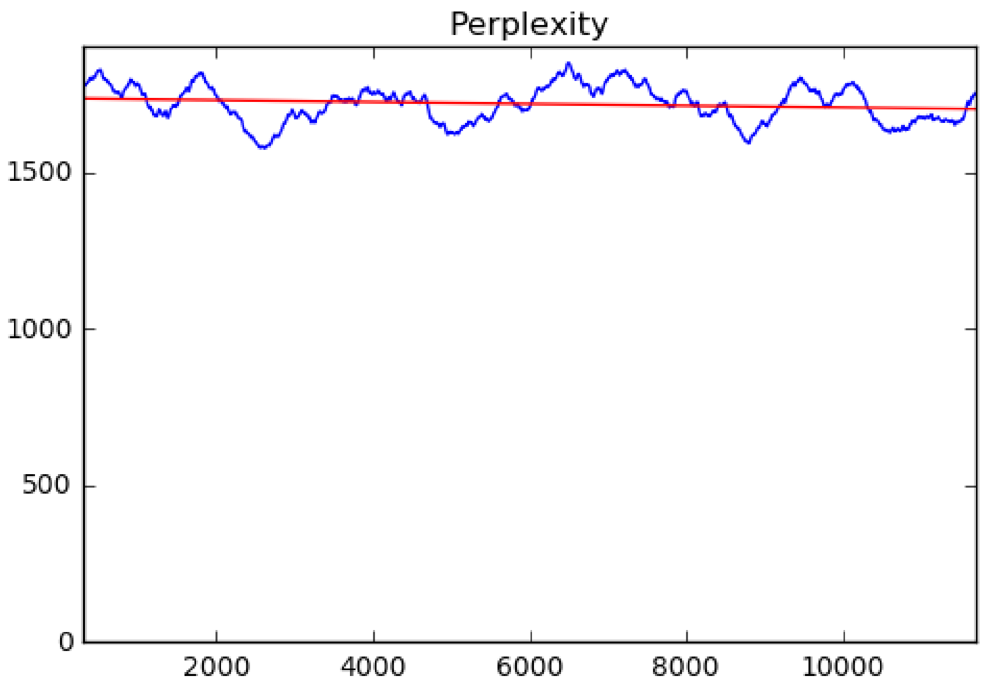
For instance, in the address for the year 2000 topic 3 is the one with the highest probability at iteration 12'000 (about 0.3 probability), but at iteration 8'000 topic 5 had the highest probability (about 0.3), whereas topic 3 had only a probability of about 0.20 at iteration 8'000.

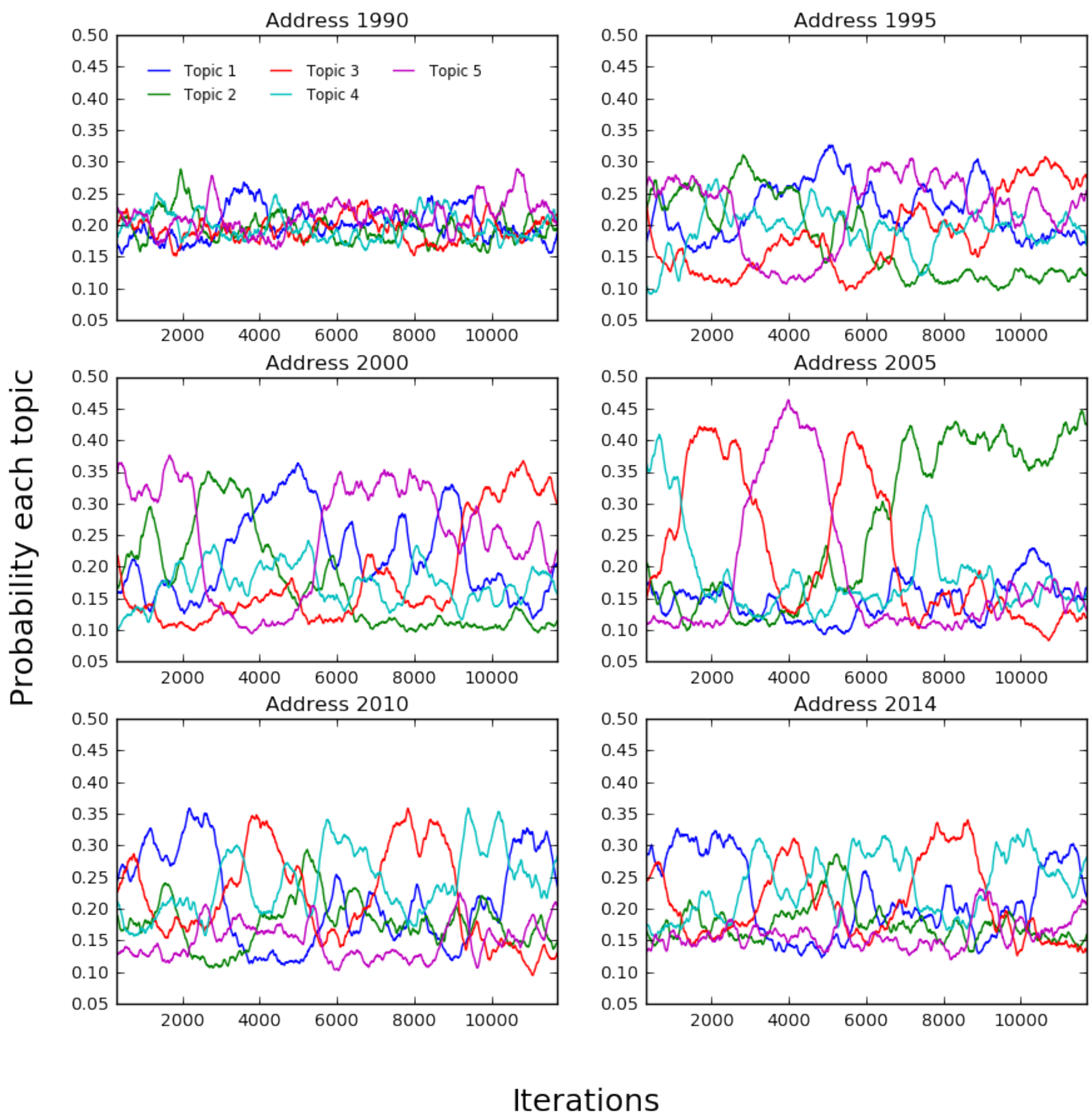
Note: if you want to plot the results of another simulation (say, with only 1000 iterations), you can run directly the second chunk of code below without loading the results of the 12'000 iterations. In that case, please adjust the variable sm_window to a smaller value (maybe 50, for 1'000 iterations) so that the moving average window is not too big for the total size of the sample.

```
In [6]: # load results from csv files saved earlier with the results of the 12'000 iterations
        perplexity = np.loadtxt("./results/perplexity.csv",delimiter=",")
        track = np.loadtxt("./results/track.csv",delimiter=",")
        track2 = np.loadtxt("./results/track2.csv",delimiter=",")
        track3 = np.loadtxt("./results/track3.csv",delimiter=",")
        track4 = np.loadtxt("./results/track4.csv",delimiter=",")
        track5 = np.loadtxt("./results/track5.csv",delimiter=",")
        track6 = np.loadtxt("./results/track6.csv",delimiter=",")
        theta_docs = np.loadtxt("./results/theta_docs.csv",delimiter=",")
        beta_terms = np.loadtxt("./results/beta_terms.csv",delimiter=",")
        TA_doc = np.loadtxt("./results/TA_doc.csv",delimiter=",")
        TA_v = np.loadtxt("./results/TA_v.csv",delimiter=",")
        TA_terms = np.loadtxt("./results/TA_terms.csv",delimiter=",")
```

```
In [17]: # Perplexity
         # smooth window: necessary as otherwise there is too much volatility and figures are
         difficult to read
         sm_window=300
         # smooth and convert to pandas
         perplexity_df = pd.DataFrame(np.convolve(perplexity,1/sm_window *
         np.ones(sm_window),'same'))
         # add trendline
         z = np.polyfit(range(iterations), perplexity, 1)
         p = np.poly1d(z)
         perplexity_df['trend'] = p(range(iterations))
         # plot
         perplexity_df.plot(legend=False, title="Perplexity",
         color=['blue','red'],xlim=(sm_window,iterations-sm_window), ylim=(0,1900))

         # Topic allocation for selected documents
         # smooth and convert to pandas
         track_df = pd.DataFrame(np.apply_along_axis(lambda m: np.convolve(m,1/sm_window *
         np.ones(sm_window),"same"),
                                             axis=0, arr=track))
         track2_df = pd.DataFrame(np.apply_along_axis(lambda m: np.convolve(m,1/sm_window *
         np.ones(sm_window),"same"),
                                             axis=0, arr=track2))
         track3_df = pd.DataFrame(np.apply_along_axis(lambda m: np.convolve(m,1/sm_window *
         np.ones(sm_window),"same"),
                                             axis=0, arr=track3))
         track4_df = pd.DataFrame(np.apply_along_axis(lambda m: np.convolve(m,1/sm_window *
         np.ones(sm_window),"same"),
                                             axis=0, arr=track4))
         track5_df = pd.DataFrame(np.apply_along_axis(lambda m: np.convolve(m,1/sm_window *
         np.ones(sm_window),"same"),
                                             axis=0, arr=track5))
         track6_df = pd.DataFrame(np.apply_along_axis(lambda m: np.convolve(m,1/sm_window *
         np.ones(sm_window),"same"),
                                             axis=0, arr=track6))

         # Create legend text
         legend = []
         for i in range(topics):
             legend.append("Topic " + str((i+1)))
         # Create a grid to fit 6 charts
         fig, ax = plt.subplots(3,2, figsize=(10,10), sharex =True, sharey=True)
         # plot first chart
         ax1 = track_df.plot(ax=ax[0,0], title="Address 1990",xlim=(sm_window,iterations-
         sm_window))
         # add legend to 1st chart
         lines, labels = ax1.get_legend_handles_labels()
         ax1.legend(loc="upper left", frameon= False,borderaxespad=1.5, labels=legend,
         ncol=3,fontsize='small')
         plt.setp(ax1.get_xticklabels(),visible=True)
```

```
# plot the rest without legend
track2_df.plot(ax=ax[0,1], legend=False, title="Address 1995",xlim=(sm_window
,iterations-sm_window))
track3_df.plot(ax=ax[1,0], legend=False, title="Address 2000", xlim=(sm_window
,iterations-sm_window))
track4_df.plot(ax=ax[1,1], legend=False, title="Address 2005",xlim=(sm_window
,iterations-sm_window))
track5_df.plot(ax=ax[2,0], legend=False, title="Address 2010",xlim=(sm_window
,iterations-sm_window))
track6_df.plot(ax=ax[2,1], legend=False, title="Address 2014",xlim=(sm_window
,iterations-sm_window))
# add x,y labels all charts
for chart in ax.flatten():
    for tk in chart.get_yticklabels():
        tk.set_visible(True)
    for tk in chart.get_xticklabels():
        tk.set_visible(True)
fig.text(0.5, 0.04, 'Iterations', ha='center', size=18)
fig.text(0.04, 0.5, 'Probability each topic', va='center', size=18, rotation='vertical')
fig.show()
```





We take the results of the Gibbs sampler (with the caveats mentioned about the convergence of the results) and produce a word cloud for each topic.

Based on the most prominent words highlighted in these diagrams, it is difficult to individuate any obvious link between each topic and external phenomena. For instance, topic 3 seems to relate to some foreign policy issues (Egypt, Syria, Kosovo), but similar terms also appear in topic 5 (e.g. Syria and Qadhafi).

Topic 1 seems to have some economic connotations (merge, shrink, jone), but the terms are rather general. Topic 2 includes some industrial/labor terms (invent, automat, overtime), whereas topic 4 is hard to relate to a particular external subject since term are rather heterodox.

From this analysis we can conclude that 5 topics are probably too few to extract distinct meaning from the State of the Union Addresses.

Note: the next chunk of code requires the package WordCloud.

```
In [97]: from wordcloud import WordCloud
```

```python
# load word frequencies for each topic
topic1 = {}
for i,term in enumerate(X.columns):
    topic1[term] = beta_terms[0,i]

topic2 = {}
for i,term in enumerate(X.columns):
    topic2[term] = beta_terms[1,i]

topic3 = {}
for i,term in enumerate(X.columns):
    topic3[term] = beta_terms[2,i]

topic4 = {}
for i,term in enumerate(X.columns):
    topic4[term] = beta_terms[3,i]

topic5 = {}
for i,term in enumerate(X.columns):
    topic5[term] = beta_terms[4,i]

# calculate wordclouds
wordcloud1 = WordCloud(relative_scaling=1,background_color='white', colormap="binary",
random_state=3).generate_from_frequencies(topic1)
wordcloud2 = WordCloud(relative_scaling=1,background_color='white', colormap="binary",
random_state=3).generate_from_frequencies(topic2)
wordcloud3 = WordCloud(relative_scaling=1,background_color='white', colormap="binary",
random_state=4).generate_from_frequencies(topic3)
wordcloud4 = WordCloud(relative_scaling=1,background_color='white', colormap="binary",
random_state=3).generate_from_frequencies(topic4)
wordcloud5 = WordCloud(relative_scaling=1,background_color='white', colormap="binary",
random_state=3).generate_from_frequencies(topic5)

#plot them
fig, ax = plt.subplots(3,2, figsize=(15,15))
ax[0,0].axis("off")
ax[0,0].set_title("Topic 1\n", size=30)
ax[0,0].imshow(wordcloud1, interpolation='bilinear')
ax[0,1].axis("off")
ax[0,1].set_title("Topic 2\n", size=30)
ax[0,1].imshow(wordcloud2, interpolation='bilinear')
ax[1,0].axis("off")
ax[1,0].set_title("Topic 3\n", size=30)
ax[1,0].imshow(wordcloud3, interpolation='bilinear')
ax[1,1].axis("off")
ax[1,1].set_title("Topic 4\n", size=30)
ax[1,1].imshow(wordcloud4, interpolation='bilinear')
ax[2,0].axis("off")
ax[2,0].set_title("Topic 5\n", size=30)
ax[2,0].imshow(wordcloud5, interpolation='bilinear')
ax[2,1].axis("off")
plt.show()
```

Topic 1

Topic 2

Topic 3

Topic 4

Topic 5

1.2.3 Exercise 2

Now we run a collapsed Gibbs sampler for the same parameter values (i.e. Dirichlet hyperparameters and K) and documents in exercise 1. We use the 'lda' package.

```
In [2]: ##2. Collapsed GIBBS SAMPLER:
        import lda

        # parameters document term matrix
        D = X.shape[0]
        V = X.shape[1]

        # parameters gibbs sampler
        topics = 5
        alpha = 10
        eta = 0.1
```

```python
iterations = 12000

model = lda.LDA(n_topics=5, n_iter=iterations, alpha = alpha, eta = eta, random_state=1)
model.fit(np.array(X))
```

```
INFO:lda:n_documents: 25
INFO:lda:vocab_size: 2070
INFO:lda:n_words: 9013
INFO:lda:n_topics: 5
INFO:lda:n_iter: 12000
INFO:lda:<0> log likelihood: -88973
INFO:lda:<10> log likelihood: -76850
INFO:lda:<20> log likelihood: -75296
INFO:lda:<30> log likelihood: -74573
INFO:lda:<40> log likelihood: -74200
INFO:lda:<50> log likelihood: -73964
INFO:lda:<60> log likelihood: -73810
INFO:lda:<70> log likelihood: -74100
INFO:lda:<80> log likelihood: -74151
INFO:lda:<90> log likelihood: -73797
INFO:lda:<100> log likelihood: -73815
INFO:lda:<110> log likelihood: -74119
INFO:lda:<120> log likelihood: -74069
INFO:lda:<130> log likelihood: -73977
INFO:lda:<140> log likelihood: -74176
INFO:lda:<150> log likelihood: -74018
INFO:lda:<160> log likelihood: -74064
INFO:lda:<170> log likelihood: -73958
INFO:lda:<180> log likelihood: -74244
INFO:lda:<190> log likelihood: -74057
INFO:lda:<200> log likelihood: -73872
INFO:lda:<210> log likelihood: -73919
INFO:lda:<220> log likelihood: -73757
INFO:lda:<230> log likelihood: -74052
INFO:lda:<240> log likelihood: -73827
INFO:lda:<250> log likelihood: -73930
INFO:lda:<260> log likelihood: -74098
INFO:lda:<270> log likelihood: -73854
INFO:lda:<280> log likelihood: -74080
INFO:lda:<290> log likelihood: -74071
INFO:lda:<300> log likelihood: -73792
INFO:lda:<310> log likelihood: -74108
INFO:lda:<320> log likelihood: -73897
INFO:lda:<330> log likelihood: -74052
INFO:lda:<340> log likelihood: -73863
INFO:lda:<350> log likelihood: -73878
INFO:lda:<360> log likelihood: -74005
INFO:lda:<370> log likelihood: -74093
INFO:lda:<380> log likelihood: -74101
INFO:lda:<390> log likelihood: -73880
INFO:lda:<400> log likelihood: -73911
INFO:lda:<410> log likelihood: -74317
INFO:lda:<420> log likelihood: -74005
INFO:lda:<430> log likelihood: -74022
INFO:lda:<440> log likelihood: -73874
INFO:lda:<450> log likelihood: -73761
INFO:lda:<460> log likelihood: -73973
INFO:lda:<470> log likelihood: -73974
INFO:lda:<480> log likelihood: -73969
```

```
INFO:lda:<490> log likelihood: -73866
INFO:lda:<500> log likelihood: -73890
INFO:lda:<510> log likelihood: -74076
INFO:lda:<520> log likelihood: -73743
INFO:lda:<530> log likelihood: -73901
INFO:lda:<540> log likelihood: -74025
INFO:lda:<550> log likelihood: -73954
INFO:lda:<560> log likelihood: -73867
INFO:lda:<570> log likelihood: -73937
INFO:lda:<580> log likelihood: -73942
INFO:lda:<590> log likelihood: -73858
INFO:lda:<600> log likelihood: -73964
INFO:lda:<610> log likelihood: -73993
INFO:lda:<620> log likelihood: -73991
INFO:lda:<630> log likelihood: -74194
INFO:lda:<640> log likelihood: -73976
INFO:lda:<650> log likelihood: -73940
INFO:lda:<660> log likelihood: -74145
INFO:lda:<670> log likelihood: -74134
INFO:lda:<680> log likelihood: -74177
INFO:lda:<690> log likelihood: -73992
INFO:lda:<700> log likelihood: -73981
INFO:lda:<710> log likelihood: -73929
INFO:lda:<720> log likelihood: -74110
INFO:lda:<730> log likelihood: -74192
INFO:lda:<740> log likelihood: -74003
INFO:lda:<750> log likelihood: -74169
INFO:lda:<760> log likelihood: -74035
INFO:lda:<770> log likelihood: -73977
INFO:lda:<780> log likelihood: -74050
INFO:lda:<790> log likelihood: -73845
INFO:lda:<800> log likelihood: -74005
INFO:lda:<810> log likelihood: -74001
INFO:lda:<820> log likelihood: -74151
INFO:lda:<830> log likelihood: -73914
INFO:lda:<840> log likelihood: -74136
INFO:lda:<850> log likelihood: -73896
INFO:lda:<860> log likelihood: -73947
INFO:lda:<870> log likelihood: -73998
INFO:lda:<880> log likelihood: -73928
INFO:lda:<890> log likelihood: -73984
INFO:lda:<900> log likelihood: -73940
INFO:lda:<910> log likelihood: -73899
INFO:lda:<920> log likelihood: -73950
INFO:lda:<930> log likelihood: -74151
INFO:lda:<940> log likelihood: -74014
INFO:lda:<950> log likelihood: -73902
INFO:lda:<960> log likelihood: -73972
INFO:lda:<970> log likelihood: -73993
INFO:lda:<980> log likelihood: -74016
INFO:lda:<990> log likelihood: -74147
INFO:lda:<1000> log likelihood: -73991
INFO:lda:<1010> log likelihood: -74177
INFO:lda:<1020> log likelihood: -73974
INFO:lda:<1030> log likelihood: -74104
INFO:lda:<1040> log likelihood: -74006
INFO:lda:<1050> log likelihood: -74153
INFO:lda:<1060> log likelihood: -74073
INFO:lda:<1070> log likelihood: -73834
```

```
INFO:lda:<1080> log likelihood: -73895
INFO:lda:<1090> log likelihood: -73798
INFO:lda:<1100> log likelihood: -73959
INFO:lda:<1110> log likelihood: -74233
INFO:lda:<1120> log likelihood: -74055
INFO:lda:<1130> log likelihood: -74042
INFO:lda:<1140> log likelihood: -73899
INFO:lda:<1150> log likelihood: -74059
INFO:lda:<1160> log likelihood: -73997
INFO:lda:<1170> log likelihood: -73885
INFO:lda:<1180> log likelihood: -73767
INFO:lda:<1190> log likelihood: -73957
INFO:lda:<1200> log likelihood: -73838
INFO:lda:<1210> log likelihood: -73796
INFO:lda:<1220> log likelihood: -73951
INFO:lda:<1230> log likelihood: -73935
INFO:lda:<1240> log likelihood: -73983
INFO:lda:<1250> log likelihood: -73959
INFO:lda:<1260> log likelihood: -73961
INFO:lda:<1270> log likelihood: -74151
INFO:lda:<1280> log likelihood: -73931
INFO:lda:<1290> log likelihood: -73936
INFO:lda:<1300> log likelihood: -74268
INFO:lda:<1310> log likelihood: -74224
INFO:lda:<1320> log likelihood: -73922
INFO:lda:<1330> log likelihood: -74112
INFO:lda:<1340> log likelihood: -74073
INFO:lda:<1350> log likelihood: -74203
INFO:lda:<1360> log likelihood: -73835
INFO:lda:<1370> log likelihood: -73961
INFO:lda:<1380> log likelihood: -74010
INFO:lda:<1390> log likelihood: -73863
INFO:lda:<1400> log likelihood: -73877
INFO:lda:<1410> log likelihood: -73925
INFO:lda:<1420> log likelihood: -73880
INFO:lda:<1430> log likelihood: -74038
INFO:lda:<1440> log likelihood: -73928
INFO:lda:<1450> log likelihood: -74007
INFO:lda:<1460> log likelihood: -73920
INFO:lda:<1470> log likelihood: -73886
INFO:lda:<1480> log likelihood: -73960
INFO:lda:<1490> log likelihood: -73962
INFO:lda:<1500> log likelihood: -73977
INFO:lda:<1510> log likelihood: -73935
INFO:lda:<1520> log likelihood: -74113
INFO:lda:<1530> log likelihood: -74062
INFO:lda:<1540> log likelihood: -73900
INFO:lda:<1550> log likelihood: -73960
INFO:lda:<1560> log likelihood: -73814
INFO:lda:<1570> log likelihood: -73941
INFO:lda:<1580> log likelihood: -74098
INFO:lda:<1590> log likelihood: -73918
INFO:lda:<1600> log likelihood: -73881
INFO:lda:<1610> log likelihood: -73795
INFO:lda:<1620> log likelihood: -74121
INFO:lda:<1630> log likelihood: -74035
INFO:lda:<1640> log likelihood: -73972
INFO:lda:<1650> log likelihood: -74028
INFO:lda:<1660> log likelihood: -73987
```

```
INFO:lda:<1670> log likelihood: -73943
INFO:lda:<1680> log likelihood: -73971
INFO:lda:<1690> log likelihood: -74125
INFO:lda:<1700> log likelihood: -73884
INFO:lda:<1710> log likelihood: -73900
INFO:lda:<1720> log likelihood: -73963
INFO:lda:<1730> log likelihood: -74045
INFO:lda:<1740> log likelihood: -73962
INFO:lda:<1750> log likelihood: -73997
INFO:lda:<1760> log likelihood: -73950
INFO:lda:<1770> log likelihood: -74059
INFO:lda:<1780> log likelihood: -73999
INFO:lda:<1790> log likelihood: -73890
INFO:lda:<1800> log likelihood: -73983
INFO:lda:<1810> log likelihood: -74052
INFO:lda:<1820> log likelihood: -73819
INFO:lda:<1830> log likelihood: -74004
INFO:lda:<1840> log likelihood: -73886
INFO:lda:<1850> log likelihood: -73883
INFO:lda:<1860> log likelihood: -73913
INFO:lda:<1870> log likelihood: -74081
INFO:lda:<1880> log likelihood: -74034
INFO:lda:<1890> log likelihood: -73861
INFO:lda:<1900> log likelihood: -73855
INFO:lda:<1910> log likelihood: -73891
INFO:lda:<1920> log likelihood: -74049
INFO:lda:<1930> log likelihood: -73959
INFO:lda:<1940> log likelihood: -73825
INFO:lda:<1950> log likelihood: -73962
INFO:lda:<1960> log likelihood: -73800
INFO:lda:<1970> log likelihood: -73951
INFO:lda:<1980> log likelihood: -73945
INFO:lda:<1990> log likelihood: -73916
INFO:lda:<2000> log likelihood: -73855
INFO:lda:<2010> log likelihood: -74048
INFO:lda:<2020> log likelihood: -73871
INFO:lda:<2030> log likelihood: -73967
INFO:lda:<2040> log likelihood: -74000
INFO:lda:<2050> log likelihood: -73967
INFO:lda:<2060> log likelihood: -73967
INFO:lda:<2070> log likelihood: -74041
INFO:lda:<2080> log likelihood: -73830
INFO:lda:<2090> log likelihood: -73855
INFO:lda:<2100> log likelihood: -74157
INFO:lda:<2110> log likelihood: -73976
INFO:lda:<2120> log likelihood: -73977
INFO:lda:<2130> log likelihood: -73851
INFO:lda:<2140> log likelihood: -73866
INFO:lda:<2150> log likelihood: -74017
INFO:lda:<2160> log likelihood: -74038
INFO:lda:<2170> log likelihood: -73833
INFO:lda:<2180> log likelihood: -74061
INFO:lda:<2190> log likelihood: -73985
INFO:lda:<2200> log likelihood: -73878
INFO:lda:<2210> log likelihood: -74044
INFO:lda:<2220> log likelihood: -73912
INFO:lda:<2230> log likelihood: -73908
INFO:lda:<2240> log likelihood: -73998
INFO:lda:<2250> log likelihood: -74088
```

```
INFO:lda:<2260> log likelihood: -74046
INFO:lda:<2270> log likelihood: -74030
INFO:lda:<2280> log likelihood: -74004
INFO:lda:<2290> log likelihood: -74115
INFO:lda:<2300> log likelihood: -73905
INFO:lda:<2310> log likelihood: -74053
INFO:lda:<2320> log likelihood: -74083
INFO:lda:<2330> log likelihood: -73999
INFO:lda:<2340> log likelihood: -74076
INFO:lda:<2350> log likelihood: -73787
INFO:lda:<2360> log likelihood: -74073
INFO:lda:<2370> log likelihood: -73905
INFO:lda:<2380> log likelihood: -73912
INFO:lda:<2390> log likelihood: -74031
INFO:lda:<2400> log likelihood: -74020
INFO:lda:<2410> log likelihood: -73949
INFO:lda:<2420> log likelihood: -73798
INFO:lda:<2430> log likelihood: -74044
INFO:lda:<2440> log likelihood: -74008
INFO:lda:<2450> log likelihood: -74105
INFO:lda:<2460> log likelihood: -74191
INFO:lda:<2470> log likelihood: -73724
INFO:lda:<2480> log likelihood: -73977
INFO:lda:<2490> log likelihood: -74017
INFO:lda:<2500> log likelihood: -73788
INFO:lda:<2510> log likelihood: -74014
INFO:lda:<2520> log likelihood: -73918
INFO:lda:<2530> log likelihood: -73885
INFO:lda:<2540> log likelihood: -73877
INFO:lda:<2550> log likelihood: -74091
INFO:lda:<2560> log likelihood: -73924
INFO:lda:<2570> log likelihood: -73782
INFO:lda:<2580> log likelihood: -73838
INFO:lda:<2590> log likelihood: -73823
INFO:lda:<2600> log likelihood: -74032
INFO:lda:<2610> log likelihood: -73907
INFO:lda:<2620> log likelihood: -73921
INFO:lda:<2630> log likelihood: -73810
INFO:lda:<2640> log likelihood: -74007
INFO:lda:<2650> log likelihood: -74066
INFO:lda:<2660> log likelihood: -73968
INFO:lda:<2670> log likelihood: -73909
INFO:lda:<2680> log likelihood: -74044
INFO:lda:<2690> log likelihood: -74037
INFO:lda:<2700> log likelihood: -73824
INFO:lda:<2710> log likelihood: -73839
INFO:lda:<2720> log likelihood: -73918
INFO:lda:<2730> log likelihood: -74024
INFO:lda:<2740> log likelihood: -73854
INFO:lda:<2750> log likelihood: -73833
INFO:lda:<2760> log likelihood: -73990
INFO:lda:<2770> log likelihood: -74191
INFO:lda:<2780> log likelihood: -74106
INFO:lda:<2790> log likelihood: -73922
INFO:lda:<2800> log likelihood: -74014
INFO:lda:<2810> log likelihood: -73900
INFO:lda:<2820> log likelihood: -73880
INFO:lda:<2830> log likelihood: -73878
INFO:lda:<2840> log likelihood: -74064
```

```
INFO:lda:<2850> log likelihood: -73858
INFO:lda:<2860> log likelihood: -73902
INFO:lda:<2870> log likelihood: -74048
INFO:lda:<2880> log likelihood: -74030
INFO:lda:<2890> log likelihood: -73753
INFO:lda:<2900> log likelihood: -73937
INFO:lda:<2910> log likelihood: -74000
INFO:lda:<2920> log likelihood: -73899
INFO:lda:<2930> log likelihood: -73994
INFO:lda:<2940> log likelihood: -73757
INFO:lda:<2950> log likelihood: -74003
INFO:lda:<2960> log likelihood: -73930
INFO:lda:<2970> log likelihood: -74003
INFO:lda:<2980> log likelihood: -74015
INFO:lda:<2990> log likelihood: -73839
INFO:lda:<3000> log likelihood: -74071
INFO:lda:<3010> log likelihood: -73950
INFO:lda:<3020> log likelihood: -74003
INFO:lda:<3030> log likelihood: -74010
INFO:lda:<3040> log likelihood: -73942
INFO:lda:<3050> log likelihood: -73831
INFO:lda:<3060> log likelihood: -73935
INFO:lda:<3070> log likelihood: -73732
INFO:lda:<3080> log likelihood: -73995
INFO:lda:<3090> log likelihood: -73889
INFO:lda:<3100> log likelihood: -73995
INFO:lda:<3110> log likelihood: -74162
INFO:lda:<3120> log likelihood: -74222
INFO:lda:<3130> log likelihood: -73966
INFO:lda:<3140> log likelihood: -73821
INFO:lda:<3150> log likelihood: -74030
INFO:lda:<3160> log likelihood: -73989
INFO:lda:<3170> log likelihood: -73959
INFO:lda:<3180> log likelihood: -73997
INFO:lda:<3190> log likelihood: -74041
INFO:lda:<3200> log likelihood: -74072
INFO:lda:<3210> log likelihood: -74133
INFO:lda:<3220> log likelihood: -73929
INFO:lda:<3230> log likelihood: -74038
INFO:lda:<3240> log likelihood: -73847
INFO:lda:<3250> log likelihood: -74126
INFO:lda:<3260> log likelihood: -73964
INFO:lda:<3270> log likelihood: -73789
INFO:lda:<3280> log likelihood: -74042
INFO:lda:<3290> log likelihood: -73921
INFO:lda:<3300> log likelihood: -74071
INFO:lda:<3310> log likelihood: -74166
INFO:lda:<3320> log likelihood: -74010
INFO:lda:<3330> log likelihood: -73869
INFO:lda:<3340> log likelihood: -74020
INFO:lda:<3350> log likelihood: -74010
INFO:lda:<3360> log likelihood: -73927
INFO:lda:<3370> log likelihood: -73912
INFO:lda:<3380> log likelihood: -74163
INFO:lda:<3390> log likelihood: -73960
INFO:lda:<3400> log likelihood: -74180
INFO:lda:<3410> log likelihood: -73872
INFO:lda:<3420> log likelihood: -74052
INFO:lda:<3430> log likelihood: -73842
```

```
INFO:lda:<3440> log likelihood: -73965
INFO:lda:<3450> log likelihood: -74049
INFO:lda:<3460> log likelihood: -74042
INFO:lda:<3470> log likelihood: -73886
INFO:lda:<3480> log likelihood: -74069
INFO:lda:<3490> log likelihood: -74051
INFO:lda:<3500> log likelihood: -74066
INFO:lda:<3510> log likelihood: -73779
INFO:lda:<3520> log likelihood: -73934
INFO:lda:<3530> log likelihood: -73906
INFO:lda:<3540> log likelihood: -73768
INFO:lda:<3550> log likelihood: -74066
INFO:lda:<3560> log likelihood: -74037
INFO:lda:<3570> log likelihood: -73911
INFO:lda:<3580> log likelihood: -74145
INFO:lda:<3590> log likelihood: -74178
INFO:lda:<3600> log likelihood: -74053
INFO:lda:<3610> log likelihood: -73890
INFO:lda:<3620> log likelihood: -73920
INFO:lda:<3630> log likelihood: -73829
INFO:lda:<3640> log likelihood: -73874
INFO:lda:<3650> log likelihood: -73888
INFO:lda:<3660> log likelihood: -74015
INFO:lda:<3670> log likelihood: -74069
INFO:lda:<3680> log likelihood: -74038
INFO:lda:<3690> log likelihood: -73859
INFO:lda:<3700> log likelihood: -73988
INFO:lda:<3710> log likelihood: -74160
INFO:lda:<3720> log likelihood: -74005
INFO:lda:<3730> log likelihood: -73899
INFO:lda:<3740> log likelihood: -74036
INFO:lda:<3750> log likelihood: -73907
INFO:lda:<3760> log likelihood: -73929
INFO:lda:<3770> log likelihood: -73981
INFO:lda:<3780> log likelihood: -73863
INFO:lda:<3790> log likelihood: -73857
INFO:lda:<3800> log likelihood: -74254
INFO:lda:<3810> log likelihood: -73963
INFO:lda:<3820> log likelihood: -74018
INFO:lda:<3830> log likelihood: -73928
INFO:lda:<3840> log likelihood: -73883
INFO:lda:<3850> log likelihood: -73980
INFO:lda:<3860> log likelihood: -73781
INFO:lda:<3870> log likelihood: -73985
INFO:lda:<3880> log likelihood: -73844
INFO:lda:<3890> log likelihood: -73947
INFO:lda:<3900> log likelihood: -73888
INFO:lda:<3910> log likelihood: -74125
INFO:lda:<3920> log likelihood: -74013
INFO:lda:<3930> log likelihood: -73796
INFO:lda:<3940> log likelihood: -73980
INFO:lda:<3950> log likelihood: -74049
INFO:lda:<3960> log likelihood: -74030
INFO:lda:<3970> log likelihood: -73964
INFO:lda:<3980> log likelihood: -74139
INFO:lda:<3990> log likelihood: -74058
INFO:lda:<4000> log likelihood: -74012
INFO:lda:<4010> log likelihood: -73898
INFO:lda:<4020> log likelihood: -74013
```

```
INFO:lda:<4030> log likelihood: -74090
INFO:lda:<4040> log likelihood: -73955
INFO:lda:<4050> log likelihood: -73949
INFO:lda:<4060> log likelihood: -73990
INFO:lda:<4070> log likelihood: -73972
INFO:lda:<4080> log likelihood: -73938
INFO:lda:<4090> log likelihood: -73921
INFO:lda:<4100> log likelihood: -74241
INFO:lda:<4110> log likelihood: -74144
INFO:lda:<4120> log likelihood: -73934
INFO:lda:<4130> log likelihood: -73991
INFO:lda:<4140> log likelihood: -74021
INFO:lda:<4150> log likelihood: -73951
INFO:lda:<4160> log likelihood: -74100
INFO:lda:<4170> log likelihood: -74276
INFO:lda:<4180> log likelihood: -73899
INFO:lda:<4190> log likelihood: -74244
INFO:lda:<4200> log likelihood: -73997
INFO:lda:<4210> log likelihood: -73992
INFO:lda:<4220> log likelihood: -74059
INFO:lda:<4230> log likelihood: -73722
INFO:lda:<4240> log likelihood: -74042
INFO:lda:<4250> log likelihood: -73944
INFO:lda:<4260> log likelihood: -73963
INFO:lda:<4270> log likelihood: -74116
INFO:lda:<4280> log likelihood: -74077
INFO:lda:<4290> log likelihood: -74082
INFO:lda:<4300> log likelihood: -74128
INFO:lda:<4310> log likelihood: -73793
INFO:lda:<4320> log likelihood: -74087
INFO:lda:<4330> log likelihood: -73896
INFO:lda:<4340> log likelihood: -73713
INFO:lda:<4350> log likelihood: -74095
INFO:lda:<4360> log likelihood: -73977
INFO:lda:<4370> log likelihood: -74163
INFO:lda:<4380> log likelihood: -73905
INFO:lda:<4390> log likelihood: -73958
INFO:lda:<4400> log likelihood: -73940
INFO:lda:<4410> log likelihood: -74196
INFO:lda:<4420> log likelihood: -74075
INFO:lda:<4430> log likelihood: -73983
INFO:lda:<4440> log likelihood: -73936
INFO:lda:<4450> log likelihood: -74047
INFO:lda:<4460> log likelihood: -73971
INFO:lda:<4470> log likelihood: -74126
INFO:lda:<4480> log likelihood: -74114
INFO:lda:<4490> log likelihood: -74120
INFO:lda:<4500> log likelihood: -74019
INFO:lda:<4510> log likelihood: -73947
INFO:lda:<4520> log likelihood: -73970
INFO:lda:<4530> log likelihood: -74082
INFO:lda:<4540> log likelihood: -73947
INFO:lda:<4550> log likelihood: -73975
INFO:lda:<4560> log likelihood: -74020
INFO:lda:<4570> log likelihood: -73936
INFO:lda:<4580> log likelihood: -73824
INFO:lda:<4590> log likelihood: -74016
INFO:lda:<4600> log likelihood: -73733
INFO:lda:<4610> log likelihood: -73893
```

```
INFO:lda:<4620> log likelihood: -74058
INFO:lda:<4630> log likelihood: -74011
INFO:lda:<4640> log likelihood: -74005
INFO:lda:<4650> log likelihood: -74165
INFO:lda:<4660> log likelihood: -74077
INFO:lda:<4670> log likelihood: -74023
INFO:lda:<4680> log likelihood: -74139
INFO:lda:<4690> log likelihood: -74240
INFO:lda:<4700> log likelihood: -74135
INFO:lda:<4710> log likelihood: -73938
INFO:lda:<4720> log likelihood: -74089
INFO:lda:<4730> log likelihood: -73791
INFO:lda:<4740> log likelihood: -74116
INFO:lda:<4750> log likelihood: -74177
INFO:lda:<4760> log likelihood: -74105
INFO:lda:<4770> log likelihood: -73858
INFO:lda:<4780> log likelihood: -74037
INFO:lda:<4790> log likelihood: -73959
INFO:lda:<4800> log likelihood: -73838
INFO:lda:<4810> log likelihood: -73852
INFO:lda:<4820> log likelihood: -74032
INFO:lda:<4830> log likelihood: -73866
INFO:lda:<4840> log likelihood: -73989
INFO:lda:<4850> log likelihood: -73923
INFO:lda:<4860> log likelihood: -74050
INFO:lda:<4870> log likelihood: -74049
INFO:lda:<4880> log likelihood: -73916
INFO:lda:<4890> log likelihood: -73950
INFO:lda:<4900> log likelihood: -73899
INFO:lda:<4910> log likelihood: -73891
INFO:lda:<4920> log likelihood: -73970
INFO:lda:<4930> log likelihood: -73905
INFO:lda:<4940> log likelihood: -74045
INFO:lda:<4950> log likelihood: -73941
INFO:lda:<4960> log likelihood: -74040
INFO:lda:<4970> log likelihood: -73821
INFO:lda:<4980> log likelihood: -73869
INFO:lda:<4990> log likelihood: -73990
INFO:lda:<5000> log likelihood: -74133
INFO:lda:<5010> log likelihood: -74026
INFO:lda:<5020> log likelihood: -73950
INFO:lda:<5030> log likelihood: -74089
INFO:lda:<5040> log likelihood: -73853
INFO:lda:<5050> log likelihood: -73794
INFO:lda:<5060> log likelihood: -74159
INFO:lda:<5070> log likelihood: -73869
INFO:lda:<5080> log likelihood: -73925
INFO:lda:<5090> log likelihood: -74114
INFO:lda:<5100> log likelihood: -74023
INFO:lda:<5110> log likelihood: -73996
INFO:lda:<5120> log likelihood: -73841
INFO:lda:<5130> log likelihood: -73839
INFO:lda:<5140> log likelihood: -74053
INFO:lda:<5150> log likelihood: -73983
INFO:lda:<5160> log likelihood: -73952
INFO:lda:<5170> log likelihood: -73904
INFO:lda:<5180> log likelihood: -73958
INFO:lda:<5190> log likelihood: -74069
INFO:lda:<5200> log likelihood: -74013
```

```
INFO:lda:<5210> log likelihood: -74092
INFO:lda:<5220> log likelihood: -73744
INFO:lda:<5230> log likelihood: -74052
INFO:lda:<5240> log likelihood: -74128
INFO:lda:<5250> log likelihood: -74172
INFO:lda:<5260> log likelihood: -73982
INFO:lda:<5270> log likelihood: -73753
INFO:lda:<5280> log likelihood: -73809
INFO:lda:<5290> log likelihood: -74067
INFO:lda:<5300> log likelihood: -73935
INFO:lda:<5310> log likelihood: -74043
INFO:lda:<5320> log likelihood: -74127
INFO:lda:<5330> log likelihood: -74130
INFO:lda:<5340> log likelihood: -73945
INFO:lda:<5350> log likelihood: -73996
INFO:lda:<5360> log likelihood: -73904
INFO:lda:<5370> log likelihood: -73850
INFO:lda:<5380> log likelihood: -74165
INFO:lda:<5390> log likelihood: -73864
INFO:lda:<5400> log likelihood: -73841
INFO:lda:<5410> log likelihood: -73961
INFO:lda:<5420> log likelihood: -73888
INFO:lda:<5430> log likelihood: -73850
INFO:lda:<5440> log likelihood: -74134
INFO:lda:<5450> log likelihood: -73843
INFO:lda:<5460> log likelihood: -73906
INFO:lda:<5470> log likelihood: -74198
INFO:lda:<5480> log likelihood: -73983
INFO:lda:<5490> log likelihood: -73987
INFO:lda:<5500> log likelihood: -73818
INFO:lda:<5510> log likelihood: -73937
INFO:lda:<5520> log likelihood: -73906
INFO:lda:<5530> log likelihood: -74056
INFO:lda:<5540> log likelihood: -74000
INFO:lda:<5550> log likelihood: -73933
INFO:lda:<5560> log likelihood: -73943
INFO:lda:<5570> log likelihood: -74116
INFO:lda:<5580> log likelihood: -74024
INFO:lda:<5590> log likelihood: -74167
INFO:lda:<5600> log likelihood: -73792
INFO:lda:<5610> log likelihood: -73990
INFO:lda:<5620> log likelihood: -73923
INFO:lda:<5630> log likelihood: -74021
INFO:lda:<5640> log likelihood: -73710
INFO:lda:<5650> log likelihood: -74038
INFO:lda:<5660> log likelihood: -74102
INFO:lda:<5670> log likelihood: -74059
INFO:lda:<5680> log likelihood: -73880
INFO:lda:<5690> log likelihood: -73955
INFO:lda:<5700> log likelihood: -73831
INFO:lda:<5710> log likelihood: -73976
INFO:lda:<5720> log likelihood: -73753
INFO:lda:<5730> log likelihood: -74059
INFO:lda:<5740> log likelihood: -74168
INFO:lda:<5750> log likelihood: -74072
INFO:lda:<5760> log likelihood: -74083
INFO:lda:<5770> log likelihood: -73954
INFO:lda:<5780> log likelihood: -73811
INFO:lda:<5790> log likelihood: -73877
```

```
INFO:lda:<5800> log likelihood: -74112
INFO:lda:<5810> log likelihood: -73826
INFO:lda:<5820> log likelihood: -73882
INFO:lda:<5830> log likelihood: -73916
INFO:lda:<5840> log likelihood: -74016
INFO:lda:<5850> log likelihood: -74030
INFO:lda:<5860> log likelihood: -74109
INFO:lda:<5870> log likelihood: -73919
INFO:lda:<5880> log likelihood: -74092
INFO:lda:<5890> log likelihood: -74174
INFO:lda:<5900> log likelihood: -74056
INFO:lda:<5910> log likelihood: -73832
INFO:lda:<5920> log likelihood: -73990
INFO:lda:<5930> log likelihood: -73894
INFO:lda:<5940> log likelihood: -73724
INFO:lda:<5950> log likelihood: -73867
INFO:lda:<5960> log likelihood: -74217
INFO:lda:<5970> log likelihood: -74234
INFO:lda:<5980> log likelihood: -74063
INFO:lda:<5990> log likelihood: -74054
INFO:lda:<6000> log likelihood: -74104
INFO:lda:<6010> log likelihood: -74192
INFO:lda:<6020> log likelihood: -73896
INFO:lda:<6030> log likelihood: -73997
INFO:lda:<6040> log likelihood: -73965
INFO:lda:<6050> log likelihood: -74067
INFO:lda:<6060> log likelihood: -73937
INFO:lda:<6070> log likelihood: -73946
INFO:lda:<6080> log likelihood: -73889
INFO:lda:<6090> log likelihood: -73915
INFO:lda:<6100> log likelihood: -74050
INFO:lda:<6110> log likelihood: -74035
INFO:lda:<6120> log likelihood: -73863
INFO:lda:<6130> log likelihood: -73751
INFO:lda:<6140> log likelihood: -74077
INFO:lda:<6150> log likelihood: -74176
INFO:lda:<6160> log likelihood: -74011
INFO:lda:<6170> log likelihood: -74028
INFO:lda:<6180> log likelihood: -73827
INFO:lda:<6190> log likelihood: -74059
INFO:lda:<6200> log likelihood: -73948
INFO:lda:<6210> log likelihood: -74189
INFO:lda:<6220> log likelihood: -73932
INFO:lda:<6230> log likelihood: -73951
INFO:lda:<6240> log likelihood: -73955
INFO:lda:<6250> log likelihood: -73930
INFO:lda:<6260> log likelihood: -73990
INFO:lda:<6270> log likelihood: -73991
INFO:lda:<6280> log likelihood: -73871
INFO:lda:<6290> log likelihood: -74007
INFO:lda:<6300> log likelihood: -74132
INFO:lda:<6310> log likelihood: -73987
INFO:lda:<6320> log likelihood: -73811
INFO:lda:<6330> log likelihood: -74004
INFO:lda:<6340> log likelihood: -73938
INFO:lda:<6350> log likelihood: -73901
INFO:lda:<6360> log likelihood: -74082
INFO:lda:<6370> log likelihood: -73837
INFO:lda:<6380> log likelihood: -73918
```

```
INFO:lda:<6390> log likelihood: -73985
INFO:lda:<6400> log likelihood: -73954
INFO:lda:<6410> log likelihood: -74081
INFO:lda:<6420> log likelihood: -73886
INFO:lda:<6430> log likelihood: -74011
INFO:lda:<6440> log likelihood: -74048
INFO:lda:<6450> log likelihood: -74044
INFO:lda:<6460> log likelihood: -74058
INFO:lda:<6470> log likelihood: -74137
INFO:lda:<6480> log likelihood: -74026
INFO:lda:<6490> log likelihood: -74113
INFO:lda:<6500> log likelihood: -73968
INFO:lda:<6510> log likelihood: -73837
INFO:lda:<6520> log likelihood: -74124
INFO:lda:<6530> log likelihood: -73899
INFO:lda:<6540> log likelihood: -73992
INFO:lda:<6550> log likelihood: -73931
INFO:lda:<6560> log likelihood: -73946
INFO:lda:<6570> log likelihood: -74028
INFO:lda:<6580> log likelihood: -74034
INFO:lda:<6590> log likelihood: -73892
INFO:lda:<6600> log likelihood: -73921
INFO:lda:<6610> log likelihood: -73939
INFO:lda:<6620> log likelihood: -73890
INFO:lda:<6630> log likelihood: -73909
INFO:lda:<6640> log likelihood: -73779
INFO:lda:<6650> log likelihood: -73891
INFO:lda:<6660> log likelihood: -73878
INFO:lda:<6670> log likelihood: -73972
INFO:lda:<6680> log likelihood: -74087
INFO:lda:<6690> log likelihood: -73964
INFO:lda:<6700> log likelihood: -73713
INFO:lda:<6710> log likelihood: -73798
INFO:lda:<6720> log likelihood: -73869
INFO:lda:<6730> log likelihood: -74140
INFO:lda:<6740> log likelihood: -73902
INFO:lda:<6750> log likelihood: -73965
INFO:lda:<6760> log likelihood: -74087
INFO:lda:<6770> log likelihood: -74000
INFO:lda:<6780> log likelihood: -74137
INFO:lda:<6790> log likelihood: -73794
INFO:lda:<6800> log likelihood: -73858
INFO:lda:<6810> log likelihood: -74104
INFO:lda:<6820> log likelihood: -73879
INFO:lda:<6830> log likelihood: -73891
INFO:lda:<6840> log likelihood: -74008
INFO:lda:<6850> log likelihood: -73817
INFO:lda:<6860> log likelihood: -73974
INFO:lda:<6870> log likelihood: -74199
INFO:lda:<6880> log likelihood: -73936
INFO:lda:<6890> log likelihood: -74126
INFO:lda:<6900> log likelihood: -73884
INFO:lda:<6910> log likelihood: -74079
INFO:lda:<6920> log likelihood: -74056
INFO:lda:<6930> log likelihood: -74029
INFO:lda:<6940> log likelihood: -74092
INFO:lda:<6950> log likelihood: -73998
INFO:lda:<6960> log likelihood: -73930
INFO:lda:<6970> log likelihood: -74015
```

```
INFO:lda:<6980> log likelihood: -74022
INFO:lda:<6990> log likelihood: -74057
INFO:lda:<7000> log likelihood: -74256
INFO:lda:<7010> log likelihood: -74056
INFO:lda:<7020> log likelihood: -74001
INFO:lda:<7030> log likelihood: -74141
INFO:lda:<7040> log likelihood: -73994
INFO:lda:<7050> log likelihood: -73968
INFO:lda:<7060> log likelihood: -74123
INFO:lda:<7070> log likelihood: -74106
INFO:lda:<7080> log likelihood: -74097
INFO:lda:<7090> log likelihood: -73929
INFO:lda:<7100> log likelihood: -73950
INFO:lda:<7110> log likelihood: -74124
INFO:lda:<7120> log likelihood: -74079
INFO:lda:<7130> log likelihood: -74064
INFO:lda:<7140> log likelihood: -74096
INFO:lda:<7150> log likelihood: -74042
INFO:lda:<7160> log likelihood: -73829
INFO:lda:<7170> log likelihood: -74033
INFO:lda:<7180> log likelihood: -74093
INFO:lda:<7190> log likelihood: -73976
INFO:lda:<7200> log likelihood: -73833
INFO:lda:<7210> log likelihood: -73954
INFO:lda:<7220> log likelihood: -74052
INFO:lda:<7230> log likelihood: -74057
INFO:lda:<7240> log likelihood: -74042
INFO:lda:<7250> log likelihood: -73926
INFO:lda:<7260> log likelihood: -73962
INFO:lda:<7270> log likelihood: -73882
INFO:lda:<7280> log likelihood: -73839
INFO:lda:<7290> log likelihood: -74076
INFO:lda:<7300> log likelihood: -73999
INFO:lda:<7310> log likelihood: -73928
INFO:lda:<7320> log likelihood: -73904
INFO:lda:<7330> log likelihood: -74015
INFO:lda:<7340> log likelihood: -73951
INFO:lda:<7350> log likelihood: -73939
INFO:lda:<7360> log likelihood: -73893
INFO:lda:<7370> log likelihood: -73826
INFO:lda:<7380> log likelihood: -73934
INFO:lda:<7390> log likelihood: -74082
INFO:lda:<7400> log likelihood: -73930
INFO:lda:<7410> log likelihood: -73898
INFO:lda:<7420> log likelihood: -73961
INFO:lda:<7430> log likelihood: -73950
INFO:lda:<7440> log likelihood: -73924
INFO:lda:<7450> log likelihood: -74111
INFO:lda:<7460> log likelihood: -73956
INFO:lda:<7470> log likelihood: -74080
INFO:lda:<7480> log likelihood: -73999
INFO:lda:<7490> log likelihood: -73820
INFO:lda:<7500> log likelihood: -74185
INFO:lda:<7510> log likelihood: -73959
INFO:lda:<7520> log likelihood: -73908
INFO:lda:<7530> log likelihood: -74044
INFO:lda:<7540> log likelihood: -74212
INFO:lda:<7550> log likelihood: -73887
INFO:lda:<7560> log likelihood: -73902
```

```
INFO:lda:<7570> log likelihood: -73996
INFO:lda:<7580> log likelihood: -73916
INFO:lda:<7590> log likelihood: -73852
INFO:lda:<7600> log likelihood: -73804
INFO:lda:<7610> log likelihood: -73853
INFO:lda:<7620> log likelihood: -73948
INFO:lda:<7630> log likelihood: -74065
INFO:lda:<7640> log likelihood: -73912
INFO:lda:<7650> log likelihood: -74170
INFO:lda:<7660> log likelihood: -73940
INFO:lda:<7670> log likelihood: -73830
INFO:lda:<7680> log likelihood: -73897
INFO:lda:<7690> log likelihood: -74053
INFO:lda:<7700> log likelihood: -73977
INFO:lda:<7710> log likelihood: -73884
INFO:lda:<7720> log likelihood: -73951
INFO:lda:<7730> log likelihood: -73956
INFO:lda:<7740> log likelihood: -73923
INFO:lda:<7750> log likelihood: -73959
INFO:lda:<7760> log likelihood: -74052
INFO:lda:<7770> log likelihood: -74143
INFO:lda:<7780> log likelihood: -74039
INFO:lda:<7790> log likelihood: -74108
INFO:lda:<7800> log likelihood: -74107
INFO:lda:<7810> log likelihood: -73917
INFO:lda:<7820> log likelihood: -73829
INFO:lda:<7830> log likelihood: -73952
INFO:lda:<7840> log likelihood: -74018
INFO:lda:<7850> log likelihood: -73875
INFO:lda:<7860> log likelihood: -74183
INFO:lda:<7870> log likelihood: -73945
INFO:lda:<7880> log likelihood: -74016
INFO:lda:<7890> log likelihood: -74003
INFO:lda:<7900> log likelihood: -73804
INFO:lda:<7910> log likelihood: -73975
INFO:lda:<7920> log likelihood: -74027
INFO:lda:<7930> log likelihood: -74068
INFO:lda:<7940> log likelihood: -74039
INFO:lda:<7950> log likelihood: -73942
INFO:lda:<7960> log likelihood: -73934
INFO:lda:<7970> log likelihood: -73819
INFO:lda:<7980> log likelihood: -74013
INFO:lda:<7990> log likelihood: -74151
INFO:lda:<8000> log likelihood: -73935
INFO:lda:<8010> log likelihood: -74070
INFO:lda:<8020> log likelihood: -74078
INFO:lda:<8030> log likelihood: -73989
INFO:lda:<8040> log likelihood: -73913
INFO:lda:<8050> log likelihood: -73945
INFO:lda:<8060> log likelihood: -74047
INFO:lda:<8070> log likelihood: -73967
INFO:lda:<8080> log likelihood: -74127
INFO:lda:<8090> log likelihood: -74122
INFO:lda:<8100> log likelihood: -73832
INFO:lda:<8110> log likelihood: -73771
INFO:lda:<8120> log likelihood: -73805
INFO:lda:<8130> log likelihood: -74123
INFO:lda:<8140> log likelihood: -73920
INFO:lda:<8150> log likelihood: -74107
```

```
INFO:lda:<8160> log likelihood: -74231
INFO:lda:<8170> log likelihood: -73855
INFO:lda:<8180> log likelihood: -74105
INFO:lda:<8190> log likelihood: -73832
INFO:lda:<8200> log likelihood: -74030
INFO:lda:<8210> log likelihood: -73924
INFO:lda:<8220> log likelihood: -74044
INFO:lda:<8230> log likelihood: -73845
INFO:lda:<8240> log likelihood: -74101
INFO:lda:<8250> log likelihood: -74005
INFO:lda:<8260> log likelihood: -73911
INFO:lda:<8270> log likelihood: -73892
INFO:lda:<8280> log likelihood: -74121
INFO:lda:<8290> log likelihood: -74083
INFO:lda:<8300> log likelihood: -74036
INFO:lda:<8310> log likelihood: -73985
INFO:lda:<8320> log likelihood: -74078
INFO:lda:<8330> log likelihood: -73974
INFO:lda:<8340> log likelihood: -74084
INFO:lda:<8350> log likelihood: -73954
INFO:lda:<8360> log likelihood: -73804
INFO:lda:<8370> log likelihood: -74041
INFO:lda:<8380> log likelihood: -74045
INFO:lda:<8390> log likelihood: -73969
INFO:lda:<8400> log likelihood: -74045
INFO:lda:<8410> log likelihood: -73884
INFO:lda:<8420> log likelihood: -74060
INFO:lda:<8430> log likelihood: -73812
INFO:lda:<8440> log likelihood: -73910
INFO:lda:<8450> log likelihood: -74051
INFO:lda:<8460> log likelihood: -73894
INFO:lda:<8470> log likelihood: -73956
INFO:lda:<8480> log likelihood: -73824
INFO:lda:<8490> log likelihood: -74053
INFO:lda:<8500> log likelihood: -74051
INFO:lda:<8510> log likelihood: -74122
INFO:lda:<8520> log likelihood: -74045
INFO:lda:<8530> log likelihood: -74087
INFO:lda:<8540> log likelihood: -74265
INFO:lda:<8550> log likelihood: -74074
INFO:lda:<8560> log likelihood: -73848
INFO:lda:<8570> log likelihood: -73993
INFO:lda:<8580> log likelihood: -74230
INFO:lda:<8590> log likelihood: -73943
INFO:lda:<8600> log likelihood: -73945
INFO:lda:<8610> log likelihood: -73893
INFO:lda:<8620> log likelihood: -73989
INFO:lda:<8630> log likelihood: -73998
INFO:lda:<8640> log likelihood: -73979
INFO:lda:<8650> log likelihood: -73953
INFO:lda:<8660> log likelihood: -73994
INFO:lda:<8670> log likelihood: -74114
INFO:lda:<8680> log likelihood: -73938
INFO:lda:<8690> log likelihood: -74075
INFO:lda:<8700> log likelihood: -73798
INFO:lda:<8710> log likelihood: -73813
INFO:lda:<8720> log likelihood: -73971
INFO:lda:<8730> log likelihood: -74056
INFO:lda:<8740> log likelihood: -73853
```

```
INFO:lda:<8750> log likelihood: -74141
INFO:lda:<8760> log likelihood: -73917
INFO:lda:<8770> log likelihood: -74027
INFO:lda:<8780> log likelihood: -74013
INFO:lda:<8790> log likelihood: -74030
INFO:lda:<8800> log likelihood: -74143
INFO:lda:<8810> log likelihood: -73975
INFO:lda:<8820> log likelihood: -73769
INFO:lda:<8830> log likelihood: -74053
INFO:lda:<8840> log likelihood: -73870
INFO:lda:<8850> log likelihood: -74167
INFO:lda:<8860> log likelihood: -74102
INFO:lda:<8870> log likelihood: -74141
INFO:lda:<8880> log likelihood: -73891
INFO:lda:<8890> log likelihood: -73977
INFO:lda:<8900> log likelihood: -73917
INFO:lda:<8910> log likelihood: -73898
INFO:lda:<8920> log likelihood: -74049
INFO:lda:<8930> log likelihood: -74099
INFO:lda:<8940> log likelihood: -73888
INFO:lda:<8950> log likelihood: -73870
INFO:lda:<8960> log likelihood: -73854
INFO:lda:<8970> log likelihood: -73999
INFO:lda:<8980> log likelihood: -74018
INFO:lda:<8990> log likelihood: -73749
INFO:lda:<9000> log likelihood: -74068
INFO:lda:<9010> log likelihood: -73812
INFO:lda:<9020> log likelihood: -73987
INFO:lda:<9030> log likelihood: -73757
INFO:lda:<9040> log likelihood: -74122
INFO:lda:<9050> log likelihood: -73872
INFO:lda:<9060> log likelihood: -73894
INFO:lda:<9070> log likelihood: -74098
INFO:lda:<9080> log likelihood: -73869
INFO:lda:<9090> log likelihood: -73885
INFO:lda:<9100> log likelihood: -73951
INFO:lda:<9110> log likelihood: -73997
INFO:lda:<9120> log likelihood: -74000
INFO:lda:<9130> log likelihood: -74039
INFO:lda:<9140> log likelihood: -73872
INFO:lda:<9150> log likelihood: -73932
INFO:lda:<9160> log likelihood: -73739
INFO:lda:<9170> log likelihood: -73921
INFO:lda:<9180> log likelihood: -74029
INFO:lda:<9190> log likelihood: -73895
INFO:lda:<9200> log likelihood: -74009
INFO:lda:<9210> log likelihood: -73882
INFO:lda:<9220> log likelihood: -73998
INFO:lda:<9230> log likelihood: -74015
INFO:lda:<9240> log likelihood: -74022
INFO:lda:<9250> log likelihood: -74174
INFO:lda:<9260> log likelihood: -74056
INFO:lda:<9270> log likelihood: -74112
INFO:lda:<9280> log likelihood: -74173
INFO:lda:<9290> log likelihood: -74098
INFO:lda:<9300> log likelihood: -73994
INFO:lda:<9310> log likelihood: -74093
INFO:lda:<9320> log likelihood: -73971
INFO:lda:<9330> log likelihood: -73852
```

```
INFO:lda:<9340> log likelihood: -73930
INFO:lda:<9350> log likelihood: -73935
INFO:lda:<9360> log likelihood: -73877
INFO:lda:<9370> log likelihood: -74072
INFO:lda:<9380> log likelihood: -74261
INFO:lda:<9390> log likelihood: -73939
INFO:lda:<9400> log likelihood: -74018
INFO:lda:<9410> log likelihood: -73980
INFO:lda:<9420> log likelihood: -73922
INFO:lda:<9430> log likelihood: -74028
INFO:lda:<9440> log likelihood: -73879
INFO:lda:<9450> log likelihood: -73909
INFO:lda:<9460> log likelihood: -73945
INFO:lda:<9470> log likelihood: -74060
INFO:lda:<9480> log likelihood: -73865
INFO:lda:<9490> log likelihood: -73933
INFO:lda:<9500> log likelihood: -74174
INFO:lda:<9510> log likelihood: -74024
INFO:lda:<9520> log likelihood: -73801
INFO:lda:<9530> log likelihood: -74102
INFO:lda:<9540> log likelihood: -73855
INFO:lda:<9550> log likelihood: -73932
INFO:lda:<9560> log likelihood: -73767
INFO:lda:<9570> log likelihood: -74086
INFO:lda:<9580> log likelihood: -74232
INFO:lda:<9590> log likelihood: -73976
INFO:lda:<9600> log likelihood: -74321
INFO:lda:<9610> log likelihood: -74037
INFO:lda:<9620> log likelihood: -74004
INFO:lda:<9630> log likelihood: -74047
INFO:lda:<9640> log likelihood: -74028
INFO:lda:<9650> log likelihood: -73931
INFO:lda:<9660> log likelihood: -74197
INFO:lda:<9670> log likelihood: -74108
INFO:lda:<9680> log likelihood: -73938
INFO:lda:<9690> log likelihood: -74087
INFO:lda:<9700> log likelihood: -74032
INFO:lda:<9710> log likelihood: -73842
INFO:lda:<9720> log likelihood: -74034
INFO:lda:<9730> log likelihood: -74151
INFO:lda:<9740> log likelihood: -73972
INFO:lda:<9750> log likelihood: -73819
INFO:lda:<9760> log likelihood: -74170
INFO:lda:<9770> log likelihood: -74011
INFO:lda:<9780> log likelihood: -74011
INFO:lda:<9790> log likelihood: -73931
INFO:lda:<9800> log likelihood: -74160
INFO:lda:<9810> log likelihood: -73835
INFO:lda:<9820> log likelihood: -73905
INFO:lda:<9830> log likelihood: -73982
INFO:lda:<9840> log likelihood: -73980
INFO:lda:<9850> log likelihood: -73892
INFO:lda:<9860> log likelihood: -73891
INFO:lda:<9870> log likelihood: -74002
INFO:lda:<9880> log likelihood: -73972
INFO:lda:<9890> log likelihood: -73915
INFO:lda:<9900> log likelihood: -73895
INFO:lda:<9910> log likelihood: -74018
INFO:lda:<9920> log likelihood: -73862
```

```
INFO:lda:<9930> log likelihood: -74055
INFO:lda:<9940> log likelihood: -74099
INFO:lda:<9950> log likelihood: -73994
INFO:lda:<9960> log likelihood: -73843
INFO:lda:<9970> log likelihood: -73898
INFO:lda:<9980> log likelihood: -74036
INFO:lda:<9990> log likelihood: -73883
INFO:lda:<10000> log likelihood: -74039
INFO:lda:<10010> log likelihood: -73872
INFO:lda:<10020> log likelihood: -74053
INFO:lda:<10030> log likelihood: -74021
INFO:lda:<10040> log likelihood: -74018
INFO:lda:<10050> log likelihood: -74020
INFO:lda:<10060> log likelihood: -74109
INFO:lda:<10070> log likelihood: -73942
INFO:lda:<10080> log likelihood: -74111
INFO:lda:<10090> log likelihood: -74015
INFO:lda:<10100> log likelihood: -73927
INFO:lda:<10110> log likelihood: -74013
INFO:lda:<10120> log likelihood: -74147
INFO:lda:<10130> log likelihood: -73889
INFO:lda:<10140> log likelihood: -74042
INFO:lda:<10150> log likelihood: -74003
INFO:lda:<10160> log likelihood: -74034
INFO:lda:<10170> log likelihood: -73717
INFO:lda:<10180> log likelihood: -73816
INFO:lda:<10190> log likelihood: -73822
INFO:lda:<10200> log likelihood: -74082
INFO:lda:<10210> log likelihood: -73969
INFO:lda:<10220> log likelihood: -74131
INFO:lda:<10230> log likelihood: -73903
INFO:lda:<10240> log likelihood: -74191
INFO:lda:<10250> log likelihood: -74021
INFO:lda:<10260> log likelihood: -73933
INFO:lda:<10270> log likelihood: -73883
INFO:lda:<10280> log likelihood: -73967
INFO:lda:<10290> log likelihood: -73806
INFO:lda:<10300> log likelihood: -73946
INFO:lda:<10310> log likelihood: -73931
INFO:lda:<10320> log likelihood: -73886
INFO:lda:<10330> log likelihood: -73901
INFO:lda:<10340> log likelihood: -73894
INFO:lda:<10350> log likelihood: -74083
INFO:lda:<10360> log likelihood: -73780
INFO:lda:<10370> log likelihood: -73930
INFO:lda:<10380> log likelihood: -73898
INFO:lda:<10390> log likelihood: -73906
INFO:lda:<10400> log likelihood: -73827
INFO:lda:<10410> log likelihood: -73941
INFO:lda:<10420> log likelihood: -73900
INFO:lda:<10430> log likelihood: -74016
INFO:lda:<10440> log likelihood: -73954
INFO:lda:<10450> log likelihood: -73949
INFO:lda:<10460> log likelihood: -74071
INFO:lda:<10470> log likelihood: -73856
INFO:lda:<10480> log likelihood: -73971
INFO:lda:<10490> log likelihood: -73793
INFO:lda:<10500> log likelihood: -74067
INFO:lda:<10510> log likelihood: -73812
```

```
INFO:lda:<10520> log likelihood: -74051
INFO:lda:<10530> log likelihood: -73811
INFO:lda:<10540> log likelihood: -73925
INFO:lda:<10550> log likelihood: -73967
INFO:lda:<10560> log likelihood: -74075
INFO:lda:<10570> log likelihood: -73896
INFO:lda:<10580> log likelihood: -74110
INFO:lda:<10590> log likelihood: -73957
INFO:lda:<10600> log likelihood: -74143
INFO:lda:<10610> log likelihood: -74006
INFO:lda:<10620> log likelihood: -73907
INFO:lda:<10630> log likelihood: -73827
INFO:lda:<10640> log likelihood: -73908
INFO:lda:<10650> log likelihood: -73882
INFO:lda:<10660> log likelihood: -73951
INFO:lda:<10670> log likelihood: -73890
INFO:lda:<10680> log likelihood: -74040
INFO:lda:<10690> log likelihood: -73980
INFO:lda:<10700> log likelihood: -73892
INFO:lda:<10710> log likelihood: -73858
INFO:lda:<10720> log likelihood: -73719
INFO:lda:<10730> log likelihood: -74136
INFO:lda:<10740> log likelihood: -73995
INFO:lda:<10750> log likelihood: -74205
INFO:lda:<10760> log likelihood: -73890
INFO:lda:<10770> log likelihood: -74006
INFO:lda:<10780> log likelihood: -73964
INFO:lda:<10790> log likelihood: -73956
INFO:lda:<10800> log likelihood: -73882
INFO:lda:<10810> log likelihood: -73986
INFO:lda:<10820> log likelihood: -73887
INFO:lda:<10830> log likelihood: -73787
INFO:lda:<10840> log likelihood: -74091
INFO:lda:<10850> log likelihood: -73989
INFO:lda:<10860> log likelihood: -73892
INFO:lda:<10870> log likelihood: -73962
INFO:lda:<10880> log likelihood: -73887
INFO:lda:<10890> log likelihood: -73906
INFO:lda:<10900> log likelihood: -73897
INFO:lda:<10910> log likelihood: -74088
INFO:lda:<10920> log likelihood: -74045
INFO:lda:<10930> log likelihood: -73692
INFO:lda:<10940> log likelihood: -74064
INFO:lda:<10950> log likelihood: -73793
INFO:lda:<10960> log likelihood: -73948
INFO:lda:<10970> log likelihood: -73856
INFO:lda:<10980> log likelihood: -74025
INFO:lda:<10990> log likelihood: -73872
INFO:lda:<11000> log likelihood: -74062
INFO:lda:<11010> log likelihood: -73853
INFO:lda:<11020> log likelihood: -73995
INFO:lda:<11030> log likelihood: -74146
INFO:lda:<11040> log likelihood: -74017
INFO:lda:<11050> log likelihood: -74098
INFO:lda:<11060> log likelihood: -74104
INFO:lda:<11070> log likelihood: -73981
INFO:lda:<11080> log likelihood: -73892
INFO:lda:<11090> log likelihood: -73981
INFO:lda:<11100> log likelihood: -73719
```

```
INFO:lda:<11110> log likelihood: -74003
INFO:lda:<11120> log likelihood: -74045
INFO:lda:<11130> log likelihood: -74000
INFO:lda:<11140> log likelihood: -73938
INFO:lda:<11150> log likelihood: -74007
INFO:lda:<11160> log likelihood: -73989
INFO:lda:<11170> log likelihood: -73987
INFO:lda:<11180> log likelihood: -73925
INFO:lda:<11190> log likelihood: -73921
INFO:lda:<11200> log likelihood: -73902
INFO:lda:<11210> log likelihood: -73972
INFO:lda:<11220> log likelihood: -74050
INFO:lda:<11230> log likelihood: -74057
INFO:lda:<11240> log likelihood: -73856
INFO:lda:<11250> log likelihood: -74016
INFO:lda:<11260> log likelihood: -74054
INFO:lda:<11270> log likelihood: -74103
INFO:lda:<11280> log likelihood: -74062
INFO:lda:<11290> log likelihood: -74025
INFO:lda:<11300> log likelihood: -74063
INFO:lda:<11310> log likelihood: -74160
INFO:lda:<11320> log likelihood: -74148
INFO:lda:<11330> log likelihood: -73978
INFO:lda:<11340> log likelihood: -74084
INFO:lda:<11350> log likelihood: -74054
INFO:lda:<11360> log likelihood: -73977
INFO:lda:<11370> log likelihood: -74190
INFO:lda:<11380> log likelihood: -73937
INFO:lda:<11390> log likelihood: -73878
INFO:lda:<11400> log likelihood: -74023
INFO:lda:<11410> log likelihood: -73862
INFO:lda:<11420> log likelihood: -73945
INFO:lda:<11430> log likelihood: -73853
INFO:lda:<11440> log likelihood: -73901
INFO:lda:<11450> log likelihood: -73984
INFO:lda:<11460> log likelihood: -73981
INFO:lda:<11470> log likelihood: -74065
INFO:lda:<11480> log likelihood: -74140
INFO:lda:<11490> log likelihood: -73930
INFO:lda:<11500> log likelihood: -74085
INFO:lda:<11510> log likelihood: -74075
INFO:lda:<11520> log likelihood: -73940
INFO:lda:<11530> log likelihood: -74066
INFO:lda:<11540> log likelihood: -73999
INFO:lda:<11550> log likelihood: -74204
INFO:lda:<11560> log likelihood: -74093
INFO:lda:<11570> log likelihood: -73924
INFO:lda:<11580> log likelihood: -74106
INFO:lda:<11590> log likelihood: -73945
INFO:lda:<11600> log likelihood: -73852
INFO:lda:<11610> log likelihood: -73800
INFO:lda:<11620> log likelihood: -73859
INFO:lda:<11630> log likelihood: -74067
INFO:lda:<11640> log likelihood: -73703
INFO:lda:<11650> log likelihood: -74054
INFO:lda:<11660> log likelihood: -73897
INFO:lda:<11670> log likelihood: -73808
INFO:lda:<11680> log likelihood: -73796
INFO:lda:<11690> log likelihood: -73875
```

```
INFO:lda:<11700> log likelihood: -74056
INFO:lda:<11710> log likelihood: -73857
INFO:lda:<11720> log likelihood: -73878
INFO:lda:<11730> log likelihood: -74169
INFO:lda:<11740> log likelihood: -73971
INFO:lda:<11750> log likelihood: -73996
INFO:lda:<11760> log likelihood: -74027
INFO:lda:<11770> log likelihood: -74004
INFO:lda:<11780> log likelihood: -73930
INFO:lda:<11790> log likelihood: -73872
INFO:lda:<11800> log likelihood: -73716
INFO:lda:<11810> log likelihood: -74001
INFO:lda:<11820> log likelihood: -73801
INFO:lda:<11830> log likelihood: -73926
INFO:lda:<11840> log likelihood: -74110
INFO:lda:<11850> log likelihood: -73945
INFO:lda:<11860> log likelihood: -73828
INFO:lda:<11870> log likelihood: -73844
INFO:lda:<11880> log likelihood: -74094
INFO:lda:<11890> log likelihood: -73868
INFO:lda:<11900> log likelihood: -73678
INFO:lda:<11910> log likelihood: -73905
INFO:lda:<11920> log likelihood: -74098
INFO:lda:<11930> log likelihood: -74095
INFO:lda:<11940> log likelihood: -73828
INFO:lda:<11950> log likelihood: -74096
INFO:lda:<11960> log likelihood: -74194
INFO:lda:<11970> log likelihood: -73856
INFO:lda:<11980> log likelihood: -74189
INFO:lda:<11990> log likelihood: -74062
INFO:lda:<11999> log likelihood: -73793
```

```
Out[2]: <lda.lda.LDA at 0x10c1729e8>
```

a) We plot the perplexity across the first 1000 sampling iterations beginning from 5 different starting values. In this case we observe the perpexitly decreases quickly during the first iterations and then it stabilizes. When compared with the perplexity of the uncollapsed Gibbs sampler (see exercise 1), we conclude the collapsed version of the sampler burns much faster.

```
In [4]: ##Perplexity
        def perplexity_iter(n_iter, X, K, alpha = np.arange(0.1,1,0.3), eta =
        np.arange(0.1,1,0.3), n_runs = 5):
            perp = np.zeros(shape = (n_runs, int(n_iter/10)))
            alpha_runs = np.zeros(n_runs)
            eta_runs = np.zeros(n_runs)
            for i in range(n_runs):
                alphai = float(np.random.choice(alpha, 1))
                etai = float(np.random.choice(eta, 1))
                alpha_runs[i]=alphai
                eta_runs[i]=etai
                model = lda.LDA(n_topics=K, n_iter=n_iter, alpha = alphai, eta = etai,
        random_state=1)
                model.fit(np.array(X))
                perp[i] = np.exp(np.negative(model.loglikelihoods_/np.sum(np.array(X))))
            return(perp, alpha_runs, eta_runs)

        n_runs = 5
        perplex, alpha_vector, eta_vector = perplexity_iter(n_iter = 1000, X = X, K = topics,
        n_runs = n_runs)
```

```
INFO:lda:n_documents: 25
INFO:lda:vocab_size: 2070
INFO:lda:n_words: 9013
INFO:lda:n_topics: 5
INFO:lda:n_iter: 1000
INFO:lda:<0> log likelihood: -85254
INFO:lda:<10> log likelihood: -75641
INFO:lda:<20> log likelihood: -72052
INFO:lda:<30> log likelihood: -71458
INFO:lda:<40> log likelihood: -70959
INFO:lda:<50> log likelihood: -70249
INFO:lda:<60> log likelihood: -70097
INFO:lda:<70> log likelihood: -70155
INFO:lda:<80> log likelihood: -70230
INFO:lda:<90> log likelihood: -70063
INFO:lda:<100> log likelihood: -70300
INFO:lda:<110> log likelihood: -70139
INFO:lda:<120> log likelihood: -70140
INFO:lda:<130> log likelihood: -70182
INFO:lda:<140> log likelihood: -70510
INFO:lda:<150> log likelihood: -70161
INFO:lda:<160> log likelihood: -70116
INFO:lda:<170> log likelihood: -70268
INFO:lda:<180> log likelihood: -70405
INFO:lda:<190> log likelihood: -70413
INFO:lda:<200> log likelihood: -70290
INFO:lda:<210> log likelihood: -70081
INFO:lda:<220> log likelihood: -70282
INFO:lda:<230> log likelihood: -70368
INFO:lda:<240> log likelihood: -70275
INFO:lda:<250> log likelihood: -70377
INFO:lda:<260> log likelihood: -70186
INFO:lda:<270> log likelihood: -70101
INFO:lda:<280> log likelihood: -70087
INFO:lda:<290> log likelihood: -70265
INFO:lda:<300> log likelihood: -70274
INFO:lda:<310> log likelihood: -70133
INFO:lda:<320> log likelihood: -70063
INFO:lda:<330> log likelihood: -70545
INFO:lda:<340> log likelihood: -70423
INFO:lda:<350> log likelihood: -70263
INFO:lda:<360> log likelihood: -70338
INFO:lda:<370> log likelihood: -70394
INFO:lda:<380> log likelihood: -70379
INFO:lda:<390> log likelihood: -70143
INFO:lda:<400> log likelihood: -70281
INFO:lda:<410> log likelihood: -70218
INFO:lda:<420> log likelihood: -70271
INFO:lda:<430> log likelihood: -70261
INFO:lda:<440> log likelihood: -70190
INFO:lda:<450> log likelihood: -70130
INFO:lda:<460> log likelihood: -70167
INFO:lda:<470> log likelihood: -70242
INFO:lda:<480> log likelihood: -70250
INFO:lda:<490> log likelihood: -70149
INFO:lda:<500> log likelihood: -70073
INFO:lda:<510> log likelihood: -70238
INFO:lda:<520> log likelihood: -70304
INFO:lda:<530> log likelihood: -70359
```

```
INFO:lda:<540> log likelihood: -70202
INFO:lda:<550> log likelihood: -70154
INFO:lda:<560> log likelihood: -70236
INFO:lda:<570> log likelihood: -70049
INFO:lda:<580> log likelihood: -70218
INFO:lda:<590> log likelihood: -70307
INFO:lda:<600> log likelihood: -70168
INFO:lda:<610> log likelihood: -70074
INFO:lda:<620> log likelihood: -70244
INFO:lda:<630> log likelihood: -70306
INFO:lda:<640> log likelihood: -70235
INFO:lda:<650> log likelihood: -70211
INFO:lda:<660> log likelihood: -70314
INFO:lda:<670> log likelihood: -70237
INFO:lda:<680> log likelihood: -70406
INFO:lda:<690> log likelihood: -70343
INFO:lda:<700> log likelihood: -70227
INFO:lda:<710> log likelihood: -70329
INFO:lda:<720> log likelihood: -70024
INFO:lda:<730> log likelihood: -70282
INFO:lda:<740> log likelihood: -70253
INFO:lda:<750> log likelihood: -70131
INFO:lda:<760> log likelihood: -70206
INFO:lda:<770> log likelihood: -70389
INFO:lda:<780> log likelihood: -70292
INFO:lda:<790> log likelihood: -70090
INFO:lda:<800> log likelihood: -70180
INFO:lda:<810> log likelihood: -70165
INFO:lda:<820> log likelihood: -70100
INFO:lda:<830> log likelihood: -70089
INFO:lda:<840> log likelihood: -70389
INFO:lda:<850> log likelihood: -70065
INFO:lda:<860> log likelihood: -70162
INFO:lda:<870> log likelihood: -70455
INFO:lda:<880> log likelihood: -70228
INFO:lda:<890> log likelihood: -70122
INFO:lda:<900> log likelihood: -70382
INFO:lda:<910> log likelihood: -69912
INFO:lda:<920> log likelihood: -69987
INFO:lda:<930> log likelihood: -70149
INFO:lda:<940> log likelihood: -69951
INFO:lda:<950> log likelihood: -70204
INFO:lda:<960> log likelihood: -70100
INFO:lda:<970> log likelihood: -70299
INFO:lda:<980> log likelihood: -70074
INFO:lda:<990> log likelihood: -70228
INFO:lda:<999> log likelihood: -70252
INFO:lda:n_documents: 25
INFO:lda:vocab_size: 2070
INFO:lda:n_words: 9013
INFO:lda:n_topics: 5
INFO:lda:n_iter: 1000
INFO:lda:<0> log likelihood: -84300
INFO:lda:<10> log likelihood: -75340
INFO:lda:<20> log likelihood: -72352
INFO:lda:<30> log likelihood: -70821
INFO:lda:<40> log likelihood: -70520
INFO:lda:<50> log likelihood: -69970
INFO:lda:<60> log likelihood: -70054
```

```
INFO:lda:<70> log likelihood: -70037
INFO:lda:<80> log likelihood: -69834
INFO:lda:<90> log likelihood: -69761
INFO:lda:<100> log likelihood: -70006
INFO:lda:<110> log likelihood: -70347
INFO:lda:<120> log likelihood: -70711
INFO:lda:<130> log likelihood: -70178
INFO:lda:<140> log likelihood: -70202
INFO:lda:<150> log likelihood: -69548
INFO:lda:<160> log likelihood: -70196
INFO:lda:<170> log likelihood: -70118
INFO:lda:<180> log likelihood: -70126
INFO:lda:<190> log likelihood: -70007
INFO:lda:<200> log likelihood: -69827
INFO:lda:<210> log likelihood: -69750
INFO:lda:<220> log likelihood: -70013
INFO:lda:<230> log likelihood: -69989
INFO:lda:<240> log likelihood: -69903
INFO:lda:<250> log likelihood: -69819
INFO:lda:<260> log likelihood: -69714
INFO:lda:<270> log likelihood: -69945
INFO:lda:<280> log likelihood: -69806
INFO:lda:<290> log likelihood: -69824
INFO:lda:<300> log likelihood: -70091
INFO:lda:<310> log likelihood: -70111
INFO:lda:<320> log likelihood: -70057
INFO:lda:<330> log likelihood: -70457
INFO:lda:<340> log likelihood: -70131
INFO:lda:<350> log likelihood: -69843
INFO:lda:<360> log likelihood: -69764
INFO:lda:<370> log likelihood: -69966
INFO:lda:<380> log likelihood: -70206
INFO:lda:<390> log likelihood: -70067
INFO:lda:<400> log likelihood: -69613
INFO:lda:<410> log likelihood: -70158
INFO:lda:<420> log likelihood: -70246
INFO:lda:<430> log likelihood: -70067
INFO:lda:<440> log likelihood: -70005
INFO:lda:<450> log likelihood: -69956
INFO:lda:<460> log likelihood: -69818
INFO:lda:<470> log likelihood: -69664
INFO:lda:<480> log likelihood: -69879
INFO:lda:<490> log likelihood: -69935
INFO:lda:<500> log likelihood: -69784
INFO:lda:<510> log likelihood: -69787
INFO:lda:<520> log likelihood: -69977
INFO:lda:<530> log likelihood: -70035
INFO:lda:<540> log likelihood: -70055
INFO:lda:<550> log likelihood: -69628
INFO:lda:<560> log likelihood: -69420
INFO:lda:<570> log likelihood: -69795
INFO:lda:<580> log likelihood: -70181
INFO:lda:<590> log likelihood: -70253
INFO:lda:<600> log likelihood: -69866
INFO:lda:<610> log likelihood: -70009
INFO:lda:<620> log likelihood: -70414
INFO:lda:<630> log likelihood: -70360
INFO:lda:<640> log likelihood: -70011
INFO:lda:<650> log likelihood: -69885
```

```
INFO:lda:<660> log likelihood: -70023
INFO:lda:<670> log likelihood: -70347
INFO:lda:<680> log likelihood: -69977
INFO:lda:<690> log likelihood: -70406
INFO:lda:<700> log likelihood: -69903
INFO:lda:<710> log likelihood: -69846
INFO:lda:<720> log likelihood: -69730
INFO:lda:<730> log likelihood: -70110
INFO:lda:<740> log likelihood: -70203
INFO:lda:<750> log likelihood: -70161
INFO:lda:<760> log likelihood: -69944
INFO:lda:<770> log likelihood: -70007
INFO:lda:<780> log likelihood: -69924
INFO:lda:<790> log likelihood: -69877
INFO:lda:<800> log likelihood: -69879
INFO:lda:<810> log likelihood: -70122
INFO:lda:<820> log likelihood: -69955
INFO:lda:<830> log likelihood: -69916
INFO:lda:<840> log likelihood: -70299
INFO:lda:<850> log likelihood: -69995
INFO:lda:<860> log likelihood: -69937
INFO:lda:<870> log likelihood: -70053
INFO:lda:<880> log likelihood: -69927
INFO:lda:<890> log likelihood: -69914
INFO:lda:<900> log likelihood: -69960
INFO:lda:<910> log likelihood: -69924
INFO:lda:<920> log likelihood: -70077
INFO:lda:<930> log likelihood: -70261
INFO:lda:<940> log likelihood: -70025
INFO:lda:<950> log likelihood: -70210
INFO:lda:<960> log likelihood: -69664
INFO:lda:<970> log likelihood: -70109
INFO:lda:<980> log likelihood: -70047
INFO:lda:<990> log likelihood: -70085
INFO:lda:<999> log likelihood: -69851
INFO:lda:n_documents: 25
INFO:lda:vocab_size: 2070
INFO:lda:n_words: 9013
INFO:lda:n_topics: 5
INFO:lda:n_iter: 1000
INFO:lda:<0> log likelihood: -89112
INFO:lda:<10> log likelihood: -76358
INFO:lda:<20> log likelihood: -74350
INFO:lda:<30> log likelihood: -73515
INFO:lda:<40> log likelihood: -73146
INFO:lda:<50> log likelihood: -72934
INFO:lda:<60> log likelihood: -72888
INFO:lda:<70> log likelihood: -72931
INFO:lda:<80> log likelihood: -72725
INFO:lda:<90> log likelihood: -72642
INFO:lda:<100> log likelihood: -72795
INFO:lda:<110> log likelihood: -72814
INFO:lda:<120> log likelihood: -72880
INFO:lda:<130> log likelihood: -72799
INFO:lda:<140> log likelihood: -72976
INFO:lda:<150> log likelihood: -72841
INFO:lda:<160> log likelihood: -73075
INFO:lda:<170> log likelihood: -72853
INFO:lda:<180> log likelihood: -72837
```

```
INFO:lda:<190> log likelihood: -73031
INFO:lda:<200> log likelihood: -72785
INFO:lda:<210> log likelihood: -72872
INFO:lda:<220> log likelihood: -72679
INFO:lda:<230> log likelihood: -72938
INFO:lda:<240> log likelihood: -72825
INFO:lda:<250> log likelihood: -72848
INFO:lda:<260> log likelihood: -72980
INFO:lda:<270> log likelihood: -72850
INFO:lda:<280> log likelihood: -72796
INFO:lda:<290> log likelihood: -72728
INFO:lda:<300> log likelihood: -72891
INFO:lda:<310> log likelihood: -72865
INFO:lda:<320> log likelihood: -72807
INFO:lda:<330> log likelihood: -72861
INFO:lda:<340> log likelihood: -72632
INFO:lda:<350> log likelihood: -72771
INFO:lda:<360> log likelihood: -72763
INFO:lda:<370> log likelihood: -72911
INFO:lda:<380> log likelihood: -72828
INFO:lda:<390> log likelihood: -72737
INFO:lda:<400> log likelihood: -72909
INFO:lda:<410> log likelihood: -73004
INFO:lda:<420> log likelihood: -73067
INFO:lda:<430> log likelihood: -72849
INFO:lda:<440> log likelihood: -72886
INFO:lda:<450> log likelihood: -72639
INFO:lda:<460> log likelihood: -72815
INFO:lda:<470> log likelihood: -72902
INFO:lda:<480> log likelihood: -72944
INFO:lda:<490> log likelihood: -72774
INFO:lda:<500> log likelihood: -72835
INFO:lda:<510> log likelihood: -73000
INFO:lda:<520> log likelihood: -72729
INFO:lda:<530> log likelihood: -72881
INFO:lda:<540> log likelihood: -73000
INFO:lda:<550> log likelihood: -72942
INFO:lda:<560> log likelihood: -72719
INFO:lda:<570> log likelihood: -72860
INFO:lda:<580> log likelihood: -72805
INFO:lda:<590> log likelihood: -72914
INFO:lda:<600> log likelihood: -72937
INFO:lda:<610> log likelihood: -72817
INFO:lda:<620> log likelihood: -72775
INFO:lda:<630> log likelihood: -72895
INFO:lda:<640> log likelihood: -72725
INFO:lda:<650> log likelihood: -72881
INFO:lda:<660> log likelihood: -73058
INFO:lda:<670> log likelihood: -72972
INFO:lda:<680> log likelihood: -73145
INFO:lda:<690> log likelihood: -73006
INFO:lda:<700> log likelihood: -72789
INFO:lda:<710> log likelihood: -72695
INFO:lda:<720> log likelihood: -72847
INFO:lda:<730> log likelihood: -72972
INFO:lda:<740> log likelihood: -72856
INFO:lda:<750> log likelihood: -72808
INFO:lda:<760> log likelihood: -72954
INFO:lda:<770> log likelihood: -73006
```

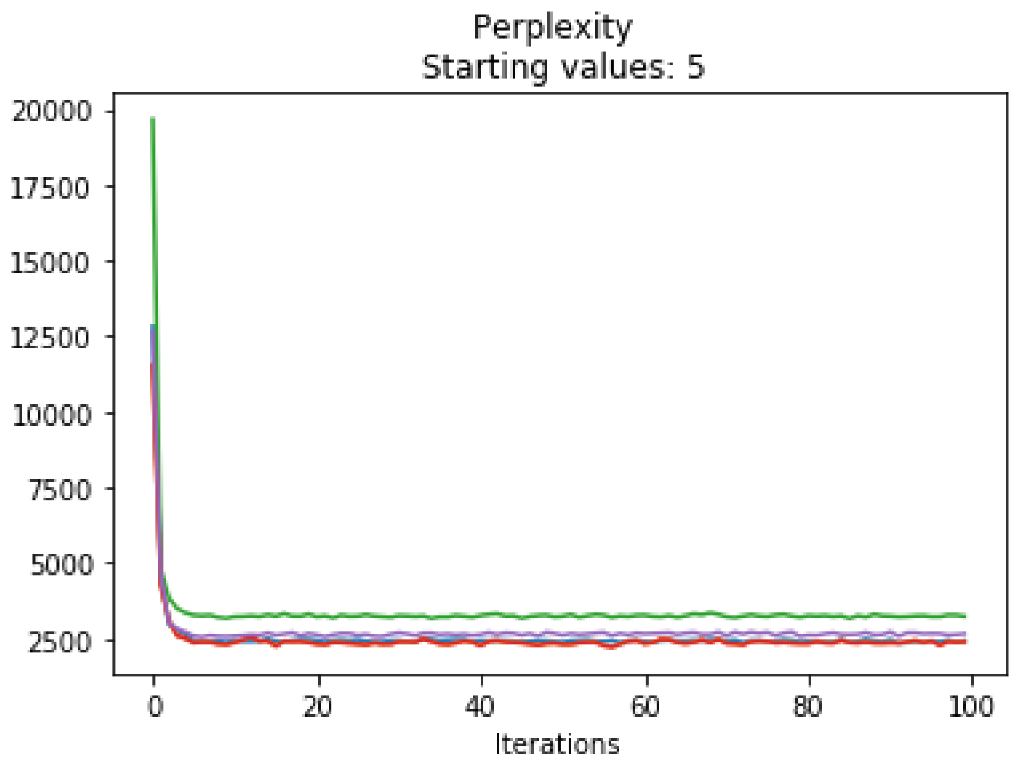
```
INFO:lda:<780> log likelihood: -72851
INFO:lda:<790> log likelihood: -72738
INFO:lda:<800> log likelihood: -72900
INFO:lda:<810> log likelihood: -72844
INFO:lda:<820> log likelihood: -72899
INFO:lda:<830> log likelihood: -72800
INFO:lda:<840> log likelihood: -72873
INFO:lda:<850> log likelihood: -72612
INFO:lda:<860> log likelihood: -72888
INFO:lda:<870> log likelihood: -72754
INFO:lda:<880> log likelihood: -73009
INFO:lda:<890> log likelihood: -72870
INFO:lda:<900> log likelihood: -72843
INFO:lda:<910> log likelihood: -72772
INFO:lda:<920> log likelihood: -72800
INFO:lda:<930> log likelihood: -72859
INFO:lda:<940> log likelihood: -72885
INFO:lda:<950> log likelihood: -72795
INFO:lda:<960> log likelihood: -72793
INFO:lda:<970> log likelihood: -72941
INFO:lda:<980> log likelihood: -72935
INFO:lda:<990> log likelihood: -72835
INFO:lda:<999> log likelihood: -72822
INFO:lda:n_documents: 25
INFO:lda:vocab_size: 2070
INFO:lda:n_words: 9013
INFO:lda:n_topics: 5
INFO:lda:n_iter: 1000
INFO:lda:<0> log likelihood: -84300
INFO:lda:<10> log likelihood: -75340
INFO:lda:<20> log likelihood: -72352
INFO:lda:<30> log likelihood: -70821
INFO:lda:<40> log likelihood: -70520
INFO:lda:<50> log likelihood: -69970
INFO:lda:<60> log likelihood: -70054
INFO:lda:<70> log likelihood: -70037
INFO:lda:<80> log likelihood: -69834
INFO:lda:<90> log likelihood: -69761
INFO:lda:<100> log likelihood: -70006
INFO:lda:<110> log likelihood: -70347
INFO:lda:<120> log likelihood: -70711
INFO:lda:<130> log likelihood: -70178
INFO:lda:<140> log likelihood: -70202
INFO:lda:<150> log likelihood: -69548
INFO:lda:<160> log likelihood: -70196
INFO:lda:<170> log likelihood: -70118
INFO:lda:<180> log likelihood: -70126
INFO:lda:<190> log likelihood: -70007
INFO:lda:<200> log likelihood: -69827
INFO:lda:<210> log likelihood: -69750
INFO:lda:<220> log likelihood: -70013
INFO:lda:<230> log likelihood: -69989
INFO:lda:<240> log likelihood: -69903
INFO:lda:<250> log likelihood: -69819
INFO:lda:<260> log likelihood: -69714
INFO:lda:<270> log likelihood: -69945
INFO:lda:<280> log likelihood: -69806
INFO:lda:<290> log likelihood: -69824
INFO:lda:<300> log likelihood: -70091
```

```
INFO:lda:<310> log likelihood: -70111
INFO:lda:<320> log likelihood: -70057
INFO:lda:<330> log likelihood: -70457
INFO:lda:<340> log likelihood: -70131
INFO:lda:<350> log likelihood: -69843
INFO:lda:<360> log likelihood: -69764
INFO:lda:<370> log likelihood: -69966
INFO:lda:<380> log likelihood: -70206
INFO:lda:<390> log likelihood: -70067
INFO:lda:<400> log likelihood: -69613
INFO:lda:<410> log likelihood: -70158
INFO:lda:<420> log likelihood: -70246
INFO:lda:<430> log likelihood: -70067
INFO:lda:<440> log likelihood: -70005
INFO:lda:<450> log likelihood: -69956
INFO:lda:<460> log likelihood: -69818
INFO:lda:<470> log likelihood: -69664
INFO:lda:<480> log likelihood: -69879
INFO:lda:<490> log likelihood: -69935
INFO:lda:<500> log likelihood: -69784
INFO:lda:<510> log likelihood: -69787
INFO:lda:<520> log likelihood: -69977
INFO:lda:<530> log likelihood: -70035
INFO:lda:<540> log likelihood: -70055
INFO:lda:<550> log likelihood: -69628
INFO:lda:<560> log likelihood: -69420
INFO:lda:<570> log likelihood: -69795
INFO:lda:<580> log likelihood: -70181
INFO:lda:<590> log likelihood: -70253
INFO:lda:<600> log likelihood: -69866
INFO:lda:<610> log likelihood: -70009
INFO:lda:<620> log likelihood: -70414
INFO:lda:<630> log likelihood: -70360
INFO:lda:<640> log likelihood: -70011
INFO:lda:<650> log likelihood: -69885
INFO:lda:<660> log likelihood: -70023
INFO:lda:<670> log likelihood: -70347
INFO:lda:<680> log likelihood: -69977
INFO:lda:<690> log likelihood: -70406
INFO:lda:<700> log likelihood: -69903
INFO:lda:<710> log likelihood: -69846
INFO:lda:<720> log likelihood: -69730
INFO:lda:<730> log likelihood: -70110
INFO:lda:<740> log likelihood: -70203
INFO:lda:<750> log likelihood: -70161
INFO:lda:<760> log likelihood: -69944
INFO:lda:<770> log likelihood: -70007
INFO:lda:<780> log likelihood: -69924
INFO:lda:<790> log likelihood: -69877
INFO:lda:<800> log likelihood: -69879
INFO:lda:<810> log likelihood: -70122
INFO:lda:<820> log likelihood: -69955
INFO:lda:<830> log likelihood: -69916
INFO:lda:<840> log likelihood: -70299
INFO:lda:<850> log likelihood: -69995
INFO:lda:<860> log likelihood: -69937
INFO:lda:<870> log likelihood: -70053
INFO:lda:<880> log likelihood: -69927
INFO:lda:<890> log likelihood: -69914
```

```
INFO:lda:<900> log likelihood: -69960
INFO:lda:<910> log likelihood: -69924
INFO:lda:<920> log likelihood: -70077
INFO:lda:<930> log likelihood: -70261
INFO:lda:<940> log likelihood: -70025
INFO:lda:<950> log likelihood: -70210
INFO:lda:<960> log likelihood: -69664
INFO:lda:<970> log likelihood: -70109
INFO:lda:<980> log likelihood: -70047
INFO:lda:<990> log likelihood: -70085
INFO:lda:<999> log likelihood: -69851
INFO:lda:n_documents: 25
INFO:lda:vocab_size: 2070
INFO:lda:n_words: 9013
INFO:lda:n_topics: 5
INFO:lda:n_iter: 1000
INFO:lda:<0> log likelihood: -85147
INFO:lda:<10> log likelihood: -75980
INFO:lda:<20> log likelihood: -72231
INFO:lda:<30> log likelihood: -71575
INFO:lda:<40> log likelihood: -71239
INFO:lda:<50> log likelihood: -70798
INFO:lda:<60> log likelihood: -70684
INFO:lda:<70> log likelihood: -70855
INFO:lda:<80> log likelihood: -70763
INFO:lda:<90> log likelihood: -70731
INFO:lda:<100> log likelihood: -70789
INFO:lda:<110> log likelihood: -70853
INFO:lda:<120> log likelihood: -70952
INFO:lda:<130> log likelihood: -70886
INFO:lda:<140> log likelihood: -71152
INFO:lda:<150> log likelihood: -70891
INFO:lda:<160> log likelihood: -71020
INFO:lda:<170> log likelihood: -71182
INFO:lda:<180> log likelihood: -70833
INFO:lda:<190> log likelihood: -71081
INFO:lda:<200> log likelihood: -70947
INFO:lda:<210> log likelihood: -70690
INFO:lda:<220> log likelihood: -70948
INFO:lda:<230> log likelihood: -71105
INFO:lda:<240> log likelihood: -71093
INFO:lda:<250> log likelihood: -70876
INFO:lda:<260> log likelihood: -70985
INFO:lda:<270> log likelihood: -70757
INFO:lda:<280> log likelihood: -70730
INFO:lda:<290> log likelihood: -71014
INFO:lda:<300> log likelihood: -71167
INFO:lda:<310> log likelihood: -71040
INFO:lda:<320> log likelihood: -70852
INFO:lda:<330> log likelihood: -71134
INFO:lda:<340> log likelihood: -71132
INFO:lda:<350> log likelihood: -70967
INFO:lda:<360> log likelihood: -71030
INFO:lda:<370> log likelihood: -70927
INFO:lda:<380> log likelihood: -71012
INFO:lda:<390> log likelihood: -70854
INFO:lda:<400> log likelihood: -70979
INFO:lda:<410> log likelihood: -71193
INFO:lda:<420> log likelihood: -70970
```

```
INFO:lda:<430> log likelihood: -71161
INFO:lda:<440> log likelihood: -71104
INFO:lda:<450> log likelihood: -70883
INFO:lda:<460> log likelihood: -71082
INFO:lda:<470> log likelihood: -70784
INFO:lda:<480> log likelihood: -71228
INFO:lda:<490> log likelihood: -71090
INFO:lda:<500> log likelihood: -70881
INFO:lda:<510> log likelihood: -71081
INFO:lda:<520> log likelihood: -70964
INFO:lda:<530> log likelihood: -71036
INFO:lda:<540> log likelihood: -71127
INFO:lda:<550> log likelihood: -71206
INFO:lda:<560> log likelihood: -70839
INFO:lda:<570> log likelihood: -71093
INFO:lda:<580> log likelihood: -70929
INFO:lda:<590> log likelihood: -71140
INFO:lda:<600> log likelihood: -71057
INFO:lda:<610> log likelihood: -71092
INFO:lda:<620> log likelihood: -71132
INFO:lda:<630> log likelihood: -71179
INFO:lda:<640> log likelihood: -70974
INFO:lda:<650> log likelihood: -71123
INFO:lda:<660> log likelihood: -71197
INFO:lda:<670> log likelihood: -70985
INFO:lda:<680> log likelihood: -70989
INFO:lda:<690> log likelihood: -71240
INFO:lda:<700> log likelihood: -70831
INFO:lda:<710> log likelihood: -71240
INFO:lda:<720> log likelihood: -71023
INFO:lda:<730> log likelihood: -71240
INFO:lda:<740> log likelihood: -71105
INFO:lda:<750> log likelihood: -71212
INFO:lda:<760> log likelihood: -70840
INFO:lda:<770> log likelihood: -71141
INFO:lda:<780> log likelihood: -71252
INFO:lda:<790> log likelihood: -70735
INFO:lda:<800> log likelihood: -70915
INFO:lda:<810> log likelihood: -70870
INFO:lda:<820> log likelihood: -71044
INFO:lda:<830> log likelihood: -71167
INFO:lda:<840> log likelihood: -70991
INFO:lda:<850> log likelihood: -70962
INFO:lda:<860> log likelihood: -70998
INFO:lda:<870> log likelihood: -71201
INFO:lda:<880> log likelihood: -70926
INFO:lda:<890> log likelihood: -70963
INFO:lda:<900> log likelihood: -71265
INFO:lda:<910> log likelihood: -70769
INFO:lda:<920> log likelihood: -71161
INFO:lda:<930> log likelihood: -71173
INFO:lda:<940> log likelihood: -71053
INFO:lda:<950> log likelihood: -71082
INFO:lda:<960> log likelihood: -71067
INFO:lda:<970> log likelihood: -70915
INFO:lda:<980> log likelihood: -70959
INFO:lda:<990> log likelihood: -71046
INFO:lda:<999> log likelihood: -71309
```

```
In [5]: for i in range(n_runs):
                plt.plot(perplex[i][0:])
        plt.xlabel('Iterations')
        plt.title('Perplexity \n Starting values: 5')
        plt.savefig('perp_nclda.png', bbox_inches='tight')
        plt.show()
```

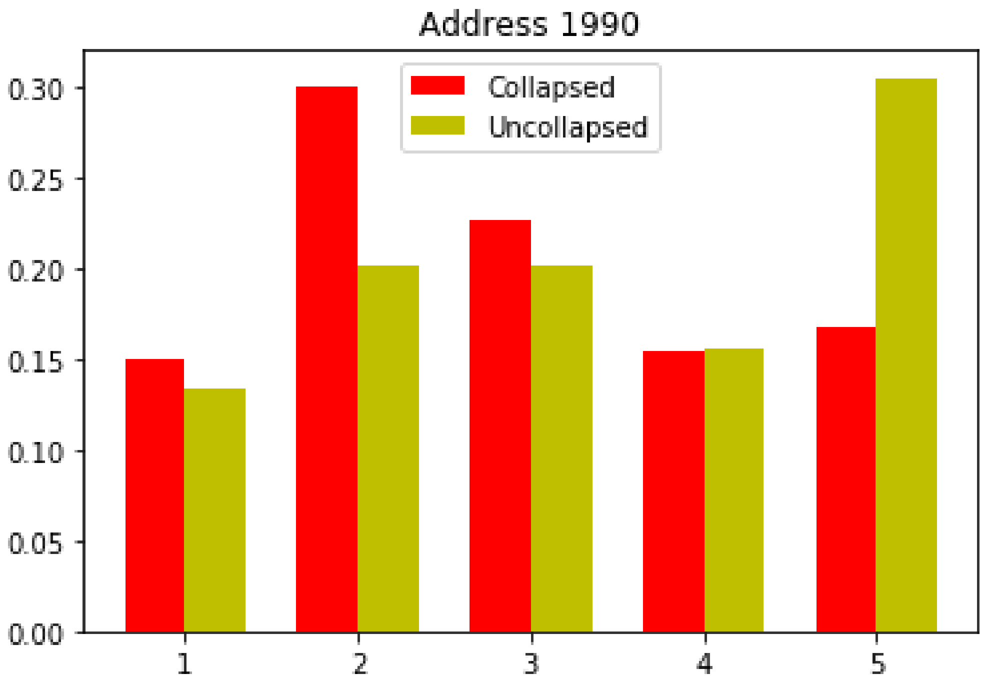


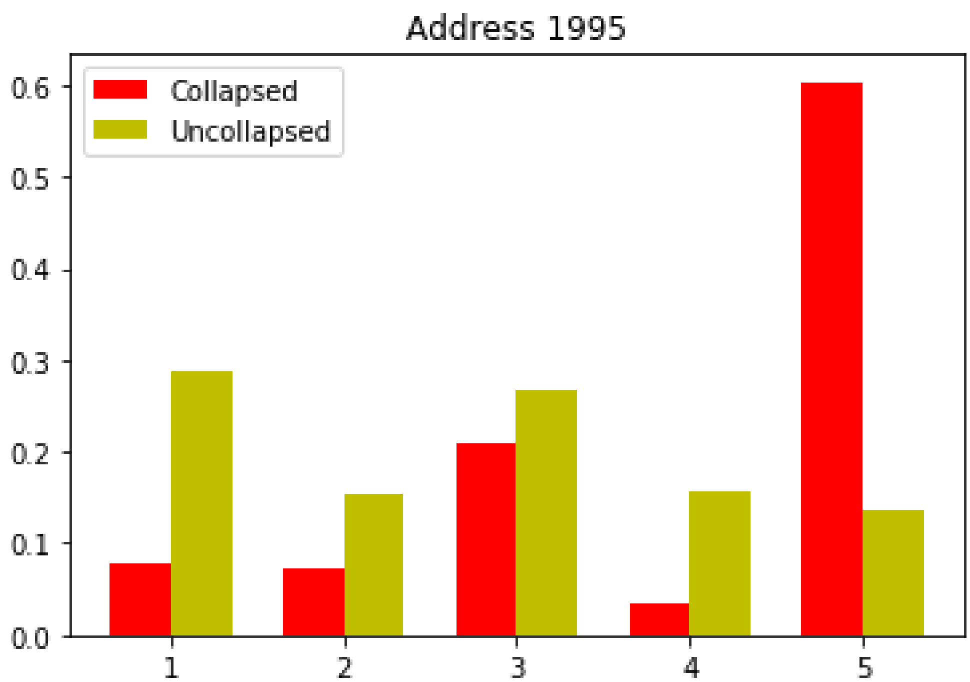
b) Now we consider estimates of the predictive distribution of θ_d for the selected documents in the previous exercise. The plots below compare the estimated topic distribution for the addresses corresponding to 1995, 2000, 2005 and 2010.

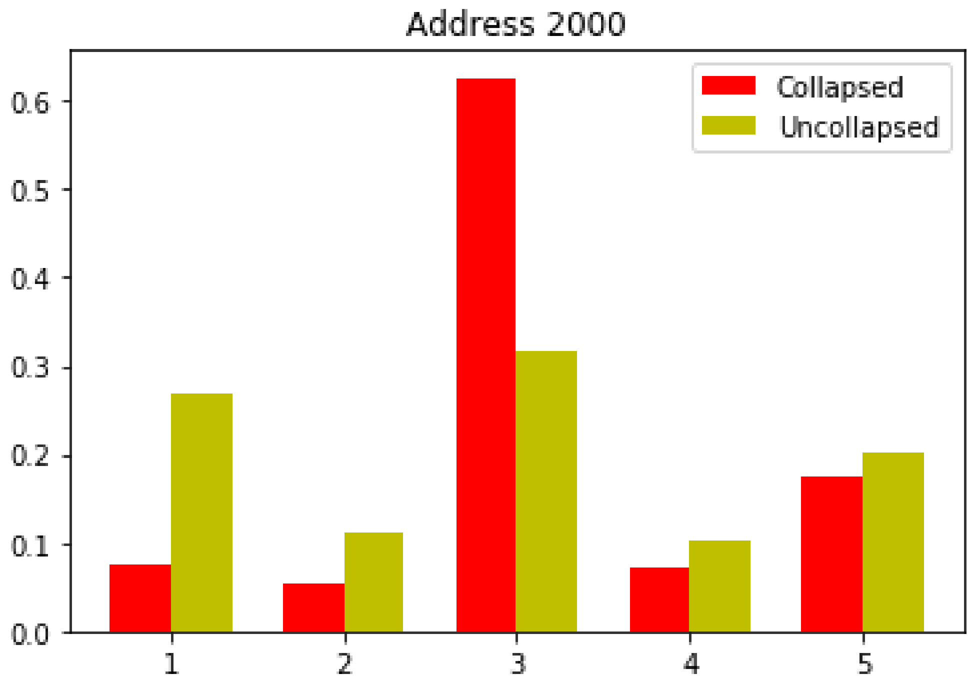
We oberserve that predictive topic distributions for the uncollapsed and collapsed samplers can be significanlty different and this could be due to the fact that the uncollapsed sampler did not converge. As we saw in the previous exercise, in the case of the uncollapsed sampler the topic distribution for the selected documents can be highly variable across iterations. However, in certain documents both sampler allow us to get the same conclusions. For example, for the 1990 address both sampler assign a similar distribution to all topics. For the 1995 address both samplers estimate a high probability to topic 3 and low probability to topics 2 and 4.

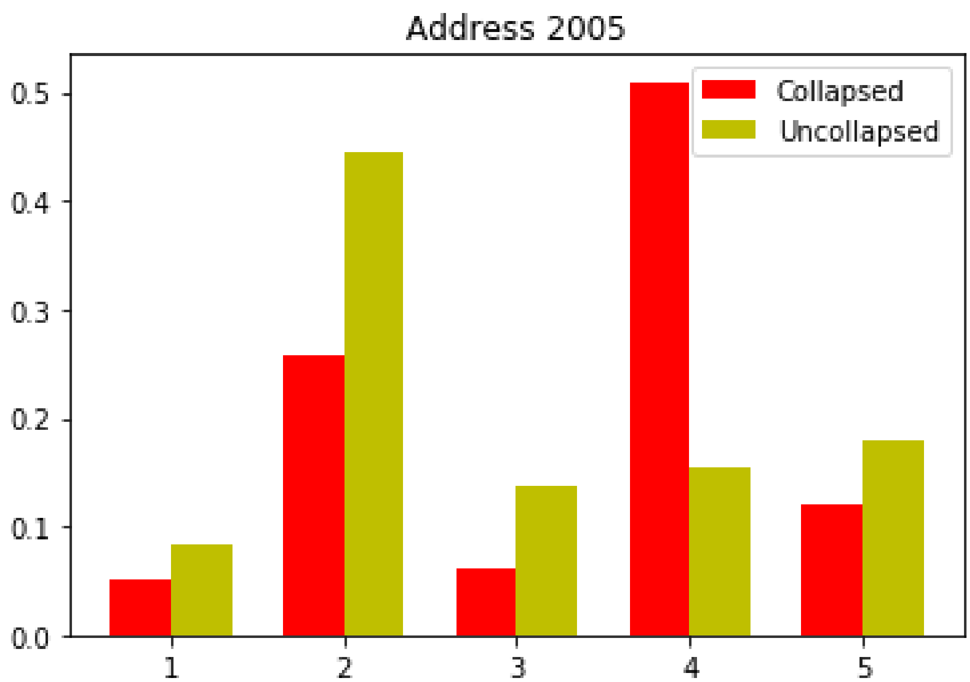
```
In [8]: track = pd.read_csv("./results/track.csv",sep=",", header = None)
        track2 = pd.read_csv("./results/track2.csv",sep=",", header = None)
        track3 = pd.read_csv("./results/track3.csv",sep=",", header = None)
        track4 = pd.read_csv("./results/track4.csv",sep=",", header = None)
        track5 = pd.read_csv("./results/track5.csv",sep=",", header = None)

        ind = np.arange(topics)  # the x locations for the groups
        width = 0.35
```

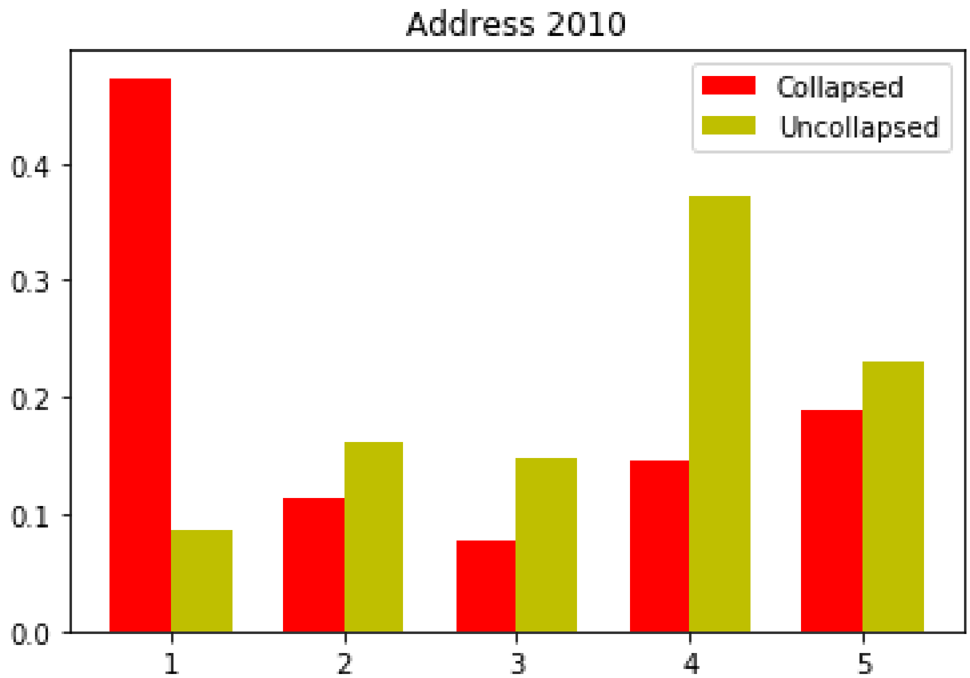
```
fig, ax = plt.subplots()
rects1 = ax.bar(ind, model.doc_topic_[0], width, color='r')
rects2 = ax.bar(ind + width, track.iloc[iterations-1], width, color='y')
ax.set_title('Address 1990')
ax.set_xticks(ind + width / 2)
ax.set_xticklabels((np.arange(topics)+1))
ax.legend((rects1[0], rects2[0]), ('Collapsed', 'Uncollapsed'))
plt.show()

fig, ax = plt.subplots()
rects1 = ax.bar(ind, model.doc_topic_[5], width, color='r')
rects2 = ax.bar(ind + width, track2.iloc[iterations-1], width, color='y')
ax.set_title('Address 1995')
ax.set_xticks(ind + width / 2)
ax.set_xticklabels((np.arange(topics)+1))
ax.legend((rects1[0], rects2[0]), ('Collapsed', 'Uncollapsed'))
plt.show()


fig, ax = plt.subplots()
rects1 = ax.bar(ind, model.doc_topic_[10], width, color='r')
rects2 = ax.bar(ind + width, track3.iloc[iterations-1], width, color='y')
ax.set_title('Address 2000')
ax.set_xticks(ind + width / 2)
ax.set_xticklabels((np.arange(topics)+1))
ax.legend((rects1[0], rects2[0]), ('Collapsed', 'Uncollapsed'))
plt.show()


fig, ax = plt.subplots()
rects1 = ax.bar(ind, model.doc_topic_[15], width, color='r')
rects2 = ax.bar(ind + width, track4.iloc[iterations-1], width, color='y')
ax.set_title('Address 2005')
ax.set_xticks(ind + width / 2)
ax.set_xticklabels((np.arange(topics)+1))
ax.legend((rects1[0], rects2[0]), ('Collapsed', 'Uncollapsed'))
plt.show()

fig, ax = plt.subplots()
rects1 = ax.bar(ind, model.doc_topic_[20], width, color='r')
rects2 = ax.bar(ind + width, track5.iloc[iterations-1], width, color='y')
ax.set_title('Address 2010')
ax.set_xticks(ind + width / 2)
ax.set_xticklabels((np.arange(topics)+1))
ax.legend((rects1[0], rects2[0]), ('Collapsed', 'Uncollapsed'))
plt.show()
```


Address 1990
Collapsed
Uncollapsed
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
2
3
4
5


Address 1995
Collapsed
Uncollapsed
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5


Address 2000
Collapsed
Uncollapsed
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
3
4
5


Address 2005
Collapsed
Uncollapsed
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
4
5



1.2.4 Exercise 3

Now we take paragraphs of state-of-the-union addresses from 1946 onwards. Each paragraph corresponds to a document and is associated with one of two political parties: Democrat or Republican. The goal is to implement a model to classify documents into one of the two political parties. In order to do so we implement a penalized logistic regression with a binary output: 1 corresponds to democrat, and 0 to republican.

In particular we implement two logistic regressions. In the first case, the paragraphs are represented as unigram counts over raw terms. Therefore, the input in the model is a document term matrix where the rows (documents) are the observations and the columns (terms) are the features. For this exercise we construct the document term matrix X considering 5000 terms. We split the sample documents in training and test data. 20% of the observations are used for testing.

We use the function `LogisticRegressionCV` which allow us to evaluate models with different penalization parameters using cross validation. Given the large number of features, we use a $L-1$ norm for the penalization so that the algorithm is allowed to set some of the coefficients to zero. Additionally, the classifier showed better out-of-sample performance with this penalty than when using a L-2 norm.

```
In [9]: ####3. Compare the classification performance
        from sklearn.feature_extraction.text import TfidfVectorizer
        from nltk.tokenize import word_tokenize

        #Consider paragraphs after 1946.
        text_data = text_raw.loc[text_raw['year']>=1946, :]

        ##1. Preprocessing of the data
        from stop_words import get_stop_words
        stop_words = get_stop_words('en')
```

```python
from nltk.stem.porter import PorterStemmer
st = PorterStemmer()

corpus = []
tokens = [] #List of all words.

for i, line in enumerate(text_data['speech']):

    #Tokenize the data:
    doc = word_tokenize(line.lower())
    #Remove non-alphabetic characters:
    doc = [tok for tok in doc if tok.isalpha()]
    #Remove stopwords using a list of your choice:
    doc = [tok for tok in doc if tok not in stop_words]
    #Stem the data using the Porter stemmer:
    doc = [st.stem(tok) for tok in doc]
    tokens.extend(doc)
    corpus.append(doc)

result = []
for i in range(0,len(corpus)):
    str1 = ' '.join(corpus[i])
    result.append(str1)

# Count Vectorizer used for words per document
from sklearn.feature_extraction.text import CountVectorizer

vectorizer = CountVectorizer(analyzer = 'word',tokenizer = word_tokenize,lowercase =
True,stop_words = 'english',max_features=5000)

X_vec = vectorizer.fit_transform(result)
feature_names = vectorizer.get_feature_names()
dense = X_vec.todense()
denselist = dense.tolist()

# Document term matrix:
X = pd.DataFrame(denselist, columns=feature_names)

# Binary output: Democrat = 1; Republican  = 0
Y = np.zeros(len(text_data))
for i in range(len(text_data)):
    if text_data.president.iloc[i] in
['Truman','Kennedy','Johnson','Carter','Clinton','Obama']:
        Y[i] = 1
    else:
        Y[i] = 0

# Divide the sample in training and test data. 20% of the observations are used for
testing.
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(np.array(X), Y,
test_size=0.2,random_state=42)

from sklearn.linear_model import LogisticRegression
from sklearn import linear_model
from sklearn import metrics
# Training Logistic regression

log_model = linear_model.LogisticRegressionCV(Cs = 50, solver = 'liblinear',
penalty='l1')
log_model.fit(X=X_train, y=Y_train)
```

```
Out[9]: LogisticRegressionCV(Cs=50, class_weight=None, cv=None, dual=False,
                   fit_intercept=True, intercept_scaling=1.0, max_iter=100,
                   multi_class='ovr', n_jobs=1, penalty='l1', random_state=None,
                   refit=True, scoring=None, solver='liblinear', tol=0.0001,
```

```
verbose=0)
```

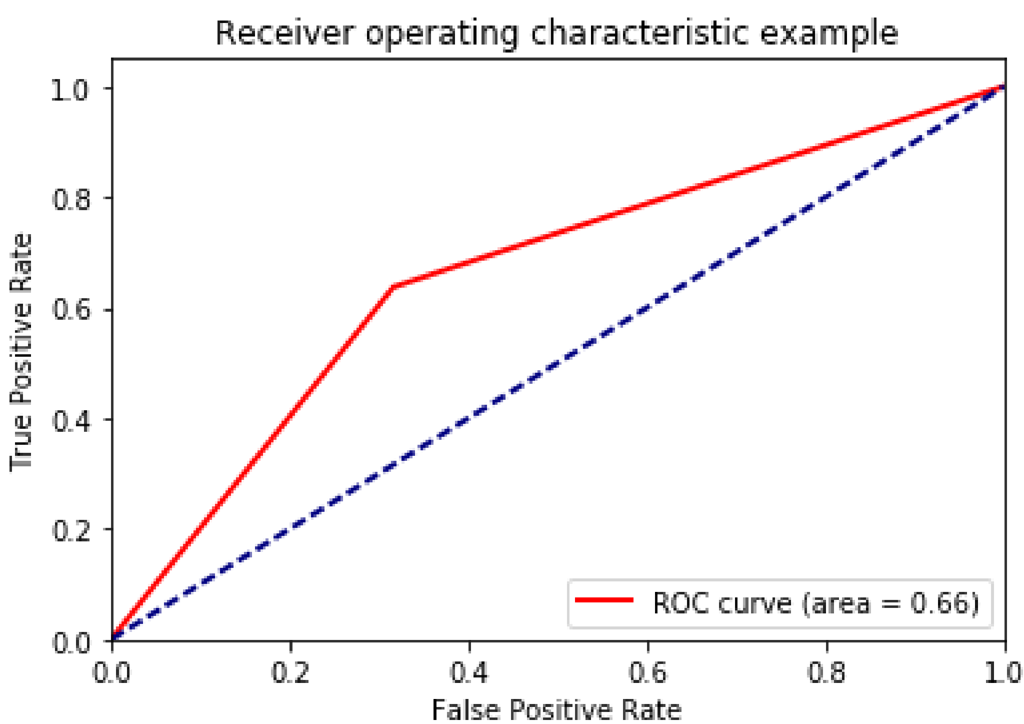
The out-of-sample performance is summarize in the following table. The *precision* is the number of true positives over the number of true positives plus the number of false positives. On the other hand, *recall* is the number of true positives over the number of true positives plus the number of false negatives.

```
In [11]: print("Logistic regression using document term matrix:\n%s\n" %
        (metrics.classification_report(Y_test,log_model.predict(X_test))))
```

```
Logistic regression using document term matrix:
             precision    recall  f1-score   support

        0.0       0.65      0.68      0.67       987
        1.0       0.67      0.64      0.66      1012

avg / total       0.66      0.66      0.66      1999
```

The next plot exhibits the ROC curve of the model.

```
In [12]: fpr, tpr, thresholds = metrics.roc_curve(Y_test, log_model.predict(X_test), pos_label=1)
         auc = metrics.auc(fpr, tpr)
         plt.figure()
         plt.plot(fpr, tpr, color = "red", lw = 2, label='ROC curve (area = %0.2f)' % auc)
         plt.plot([0, 1], [0, 1], color='navy', lw = 2, linestyle='--')
         plt.xlim([0.0, 1.0])
         plt.ylim([0.0, 1.05])
         plt.xlabel('False Positive Rate')
         plt.ylabel('True Positive Rate')
         plt.title('Receiver operating characteristic example')
         plt.legend(loc="lower right")
         plt.show()
```



In the second case, we perform classification using topic shares. The input in the regression is the matrix θ which is the distribution of documents over topics in the LDA model. We estimate θ after running a collapsed Gibbs sampler on the document term matrix X. For this exercise we define 100 topics.

```
In [13]: model = lda.LDA(n_topics=100, n_iter=1000, alpha = 0.1, eta = 0.1, random_state=1)
         model.fit(np.array(X))

         theta = model.doc_topic_
         X_train, X_test, Y_train, Y_test = train_test_split(theta, Y,
         test_size=0.2,random_state=42)

         log_model_topic = linear_model.LogisticRegressionCV(Cs = 50, solver = 'liblinear',
         penalty='l1')
         log_model_topic.fit(X=X_train, y=Y_train)
```

```
INFO:lda:n_documents: 9994
INFO:lda:vocab_size: 5000
INFO:lda:n_words: 258137
INFO:lda:n_topics: 100
INFO:lda:n_iter: 1000
WARNING:lda:all zero row in document-term matrix found
INFO:lda:<0> log likelihood: -3433943
INFO:lda:<10> log likelihood: -2460190
INFO:lda:<20> log likelihood: -2319227
INFO:lda:<30> log likelihood: -2267390
INFO:lda:<40> log likelihood: -2248245
INFO:lda:<50> log likelihood: -2234464
INFO:lda:<60> log likelihood: -2227522
INFO:lda:<70> log likelihood: -2224672
INFO:lda:<80> log likelihood: -2220740
INFO:lda:<90> log likelihood: -2218734
INFO:lda:<100> log likelihood: -2217713
INFO:lda:<110> log likelihood: -2215880
INFO:lda:<120> log likelihood: -2214532
INFO:lda:<130> log likelihood: -2215005
INFO:lda:<140> log likelihood: -2211035
INFO:lda:<150> log likelihood: -2211330
INFO:lda:<160> log likelihood: -2208880
INFO:lda:<170> log likelihood: -2209183
INFO:lda:<180> log likelihood: -2210251
INFO:lda:<190> log likelihood: -2208098
INFO:lda:<200> log likelihood: -2208744
INFO:lda:<210> log likelihood: -2208473
INFO:lda:<220> log likelihood: -2207955
INFO:lda:<230> log likelihood: -2208450
INFO:lda:<240> log likelihood: -2206578
INFO:lda:<250> log likelihood: -2206193
INFO:lda:<260> log likelihood: -2208227
INFO:lda:<270> log likelihood: -2207914
INFO:lda:<280> log likelihood: -2205032
INFO:lda:<290> log likelihood: -2204800
INFO:lda:<300> log likelihood: -2206959
INFO:lda:<310> log likelihood: -2206283
INFO:lda:<320> log likelihood: -2203148
INFO:lda:<330> log likelihood: -2206207
INFO:lda:<340> log likelihood: -2206361
```

```
INFO:lda:<350> log likelihood: -2204340
INFO:lda:<360> log likelihood: -2204969
INFO:lda:<370> log likelihood: -2204562
INFO:lda:<380> log likelihood: -2205770
INFO:lda:<390> log likelihood: -2205522
INFO:lda:<400> log likelihood: -2204151
INFO:lda:<410> log likelihood: -2206408
INFO:lda:<420> log likelihood: -2206104
INFO:lda:<430> log likelihood: -2206085
INFO:lda:<440> log likelihood: -2203980
INFO:lda:<450> log likelihood: -2204582
INFO:lda:<460> log likelihood: -2205445
INFO:lda:<470> log likelihood: -2206371
INFO:lda:<480> log likelihood: -2204345
INFO:lda:<490> log likelihood: -2205520
INFO:lda:<500> log likelihood: -2203742
INFO:lda:<510> log likelihood: -2205532
INFO:lda:<520> log likelihood: -2203376
INFO:lda:<530> log likelihood: -2202390
INFO:lda:<540> log likelihood: -2202926
INFO:lda:<550> log likelihood: -2203241
INFO:lda:<560> log likelihood: -2203949
INFO:lda:<570> log likelihood: -2203444
INFO:lda:<580> log likelihood: -2203946
INFO:lda:<590> log likelihood: -2204776
INFO:lda:<600> log likelihood: -2203846
INFO:lda:<610> log likelihood: -2203692
INFO:lda:<620> log likelihood: -2202351
INFO:lda:<630> log likelihood: -2204718
INFO:lda:<640> log likelihood: -2202117
INFO:lda:<650> log likelihood: -2204920
INFO:lda:<660> log likelihood: -2203381
INFO:lda:<670> log likelihood: -2203167
INFO:lda:<680> log likelihood: -2202713
INFO:lda:<690> log likelihood: -2201081
INFO:lda:<700> log likelihood: -2201795
INFO:lda:<710> log likelihood: -2202752
INFO:lda:<720> log likelihood: -2203563
INFO:lda:<730> log likelihood: -2202969
INFO:lda:<740> log likelihood: -2203335
INFO:lda:<750> log likelihood: -2201720
INFO:lda:<760> log likelihood: -2204263
INFO:lda:<770> log likelihood: -2203397
INFO:lda:<780> log likelihood: -2203706
INFO:lda:<790> log likelihood: -2203002
INFO:lda:<800> log likelihood: -2203311
INFO:lda:<810> log likelihood: -2203671
INFO:lda:<820> log likelihood: -2203110
INFO:lda:<830> log likelihood: -2201092
INFO:lda:<840> log likelihood: -2201586
INFO:lda:<850> log likelihood: -2202369
INFO:lda:<860> log likelihood: -2202862
INFO:lda:<870> log likelihood: -2202719
INFO:lda:<880> log likelihood: -2203433
INFO:lda:<890> log likelihood: -2204231
INFO:lda:<900> log likelihood: -2203130
INFO:lda:<910> log likelihood: -2202018
INFO:lda:<920> log likelihood: -2200837
INFO:lda:<930> log likelihood: -2201909
```

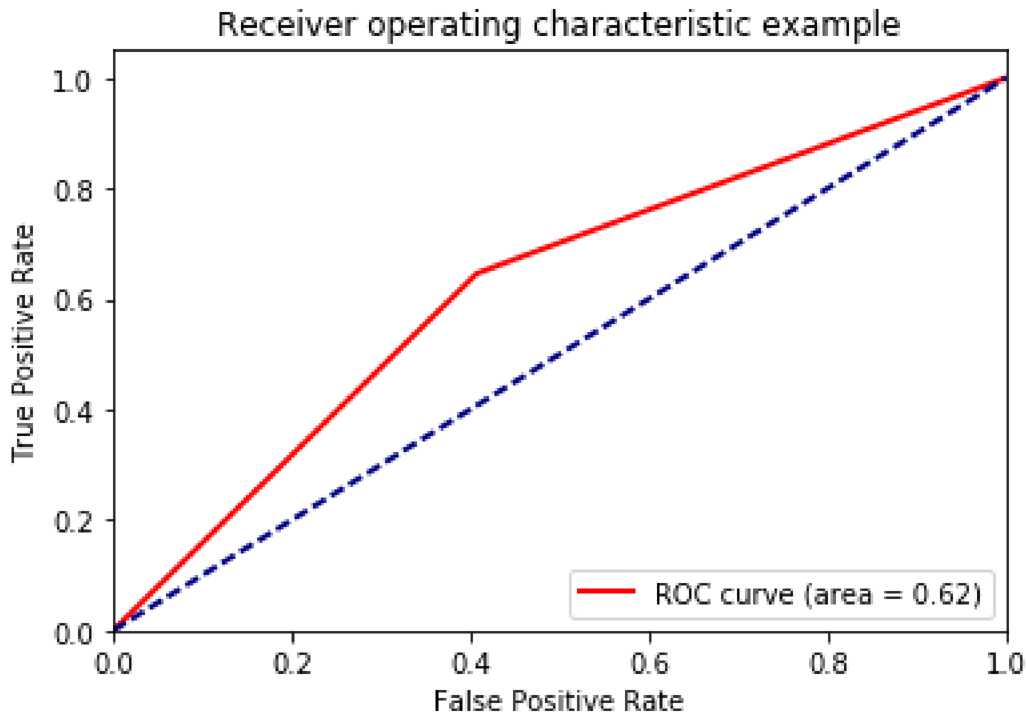
```
INFO:lda:<940> log likelihood: -2201390
INFO:lda:<950> log likelihood: -2201184
INFO:lda:<960> log likelihood: -2201275
INFO:lda:<970> log likelihood: -2202463
INFO:lda:<980> log likelihood: -2200662
INFO:lda:<990> log likelihood: -2201967
INFO:lda:<999> log likelihood: -2200530
```

```
Out[13]: LogisticRegressionCV(Cs=50, class_weight=None, cv=None, dual=False,
                    fit_intercept=True, intercept_scaling=1.0, max_iter=100,
                    multi_class='ovr', n_jobs=1, penalty='l1', random_state=None,
                    refit=True, scoring=None, solver='liblinear', tol=0.0001,
                    verbose=0)
```

The out-of-sample performance is summarize in the following table. The plot exhibits the ROC curve for the second model.

```
In [14]: print("Logistic regression using topic shares:\n%s\n" %
         (metrics.classification_report(Y_test,log_model_topic.predict(X_test))))

         fpr, tpr, thresholds = metrics.roc_curve(Y_test, log_model_topic.predict(X_test),
         pos_label=1)
         auc = metrics.auc(fpr, tpr)
         plt.figure()
         plt.plot(fpr, tpr, color = "red", lw = 2, label='ROC curve (area = %0.2f)' % auc)
         plt.plot([0, 1], [0, 1], color='navy', lw = 2, linestyle='--')
         plt.xlim([0.0, 1.0])
         plt.ylim([0.0, 1.05])
         plt.xlabel('False Positive Rate')
         plt.ylabel('True Positive Rate')
         plt.title('Receiver operating characteristic example')
         plt.legend(loc="lower right")
         plt.show()
```

```
Logistic regression using topic shares:
             precision    recall  f1-score   support

        0.0       0.62      0.59      0.61       987
        1.0       0.62      0.65      0.63      1012

avg / total       0.62      0.62      0.62      1999
```



Given the ROC curve associated with each model, and the precision and recall measures, we conclude that the regression on the raw term counts presents better results than the regression on topic shares. However, the difference is out-of-sample performance is not significantly different, which tell us that the dimension reduction properly captures the relation between the documents and the political party.