HW 5: Clustering and Topic Modeling

> Each assignment needs to be completed independently. Never ever copy others' work (even with minor modification, e.g. changing variable names). Anti-Plagiarism software will be used to check all submissions.

In this assignment, you'll need to use the following dataset:

- text_train.json: This file contains a list of documents. It's used for training models
- text_test.json: This file contains a list of documents and their ground-truth labels. It's used for testing performance. This file is in the format shown below. Note, a document may have multiple labels.

Note: due to randomness, every time you run your clustering models, you may get different results. To ease the grading process, once you get satisfactory results, please save your notebook as a pdf file (Jupyter notebook menu File -> Print -> Save as pdf), and submit this pdf along with your .py code.

In [9]:
```
import pandas as pd

from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.cluster import KMeansClusterer
import numpy as np
import pandas as pd
from sklearn import metrics
from nltk.corpus import stopwords
from nltk.cluster import KMeansClusterer, \
cosine_distance
from nltk.cluster import KMeansClusterer
from sklearn.mixture import GaussianMixture
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"
```

In [4]:
```
train_data = pd.read_csv("hw5_train.csv")
train_data.head()

test_data = pd.read_csv("hw5_test.csv")
test_data.head()
```

Out[4]:

	text
0	blm in wyo to begin deciding on backlogged lea...
1	report amtrak loss comes to per passenger u s ...
2	medicare key in races washington an upset vict...
3	sunnyvale bicyclist dies of injuries suffered ...
4	mozambique upbeat on debt crisis investors not...

Out[4]:

	text	T1	T2	T3
0	child asylum seekers targeted in home office b...	0	1	0
1	obama acknowledges economic stress not so long...	0	1	0
2	help not soonbyline by ted ralltime fri jul pm...	0	1	0
3	un pakistan flood misery exceeds tsunami haiti...	1	0	0
4	new home sales plunge new home sales plunge we...	0	1	0

Q1: K-Mean Clustering (5 points)

Define a function `cluster_kmean(train_data, test_data, num_clusters, min_df = 1, stopwords = None, metric = 'cosine')` as follows:

- Take two dataframes as inputs: `train_data` is the dataframe loaded from `hw5_train.csv` , and `test_data` is the dataframe loaded from `hw5_test.csv`
- Use **KMeans** to cluster documents in `train_data` into 3 clusters by the distance metric specified. Tune the following parameters carefully:
  - `min_df` and `stopword` options in generating TFIDF matrix. You may need to remove corpus-specific stopwords in addition to the standard stopwords.
  - distance metric: `cosine` or `Euclidean` distance
  - sufficient iterations with different initial centroids to make sure clustering converges
- Test the clustering model performance using `test_data` :
  - Predict the cluster ID for each document in `test_data` .
  - Apply `majority vote` rule to dynamically map each cluster to a ground-truth label in `test_data` .
    - Note a small percentage of documents have multiple labels. For these cases, you can randomly pick a label during the match
    - Be sure `not to hardcode the mapping` , because a cluster may corrspond to a different topic in each run. (hint: if you use pandas, look for `idxmax` function)
  - Calculate `precision/recall/f-score` for each label. Your best F1 score on the test dataset should be around `80%` .
- Assign a meaninful name to each cluster based on the `top keywords` in each cluster. You can print out the keywords and write the cluster names as markdown comments.
- This function has no return. Print out confusion matrix, precision/recall/f-score.

Analysis:

- Comparing the clustering with cosine distance and that with Euclidean distance, do you notice any difference? Which metric works better here?
- How would the stopwords and min_df options affect your clustering results?

In [5]:
```
# adding label col in test data
def get_label(row):
    if row['T1'] == 1:
        return 'T1'
    elif row['T2'] == 1:
        return 'T2'
    elif row['T3'] == 1:
        return 'T3'
    else:
        return None

test_data['label'] = test_data.apply(lambda row: get_label(row), axis=
test_data.head()
```

Out[5]:

	text	T1	T2	T3	label
0	child asylum seekers targeted in home office b...	0	1	0	T2
1	obama acknowledges economic stress not so long...	0	1	0	T2
2	help not soonbyline by ted ralltime fri jul pm...	0	1	0	T2
3	un pakistan flood misery exceeds tsunami haiti...	1	0	0	T1
4	new home sales plunge new home sales plunge we...	0	1	0	T2

In [74]:
```
def cluster_kmean(train_data, test_data, num_clusters, min_df=5, stopw

    tfidf_vect = TfidfVectorizer(stop_words="english", min_df=5)
    dtm = tfidf_vect.fit_transform(train_data["text"])


    # initialize clustering model

    clusterer = KMeansClusterer(num_clusters, cosine_distance, repeats

    # samples are assigned to cluster labels
    clusters = clusterer.cluster(dtm.toarray(), assign_clusters=True)

    centroids = np.array(clusterer.means())

    # argsort sort the matrix in ascending order
    # and return locations of features before sorting
    # [:,::-1] reverse the order
    sorted_centroids = centroids.argsort()[:, ::-1]

    # The mapping between feature (word)
    # index and feature (word) can be obtained by
    # the vectorizer's function get_feature_names()
    voc_lookup = tfidf_vect.get_feature_names()
```

```
for i in range(num_clusters):
    # get words with top 20 tf-idf weight in the centroid
    top_words = [voc_lookup[word_index] for word_index in sorted_c
    print("Cluster %d:\n %s " % (i, "; ".join(top_words)))

test_dtm = tfidf_vect.transform(test_data["text"])

predicted = [clusterer.classify(v) for v in test_dtm.toarray()]
confusion_df = pd.DataFrame(list(zip(test_data["label"].values, pr

print()

# generate crosstab between clusters and true labels
print(pd.crosstab(index=confusion_df.cluster, columns=confusion_df

cluster_dict = {0:'T1', 1:"T2", 2:'T3'}

# Map true label to cluster id
predicted_target = [cluster_dict[i] for i in predicted]

print()

print(metrics.classification_report(test_data["label"], predicted_
```

In [75]:
```
num_clusters = 3
cluster_kmean(train_data, test_data, num_clusters, min_df=5, metric='c
```

```
/Users/yashitavajpayee/opt/anaconda3/lib/python3.9/site-packages/sklearn/utils/deprecation.py:87: FutureWarning: Function get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)

Cluster 0:
 said; crash; bus; plane; police; passengers; train; car; cruise; airlines; accident; flight; road; says; driver; airport; people; traffic; injured; air
Cluster 1:
 said; oil; bp; people; japan; water; spill; disaster; gulf; earthquake; nuclear; pakistan; tsunami; quake; floods; relief; million; government; plant; coast
Cluster 2:
 percent; tax; said; year; obama; economy; rate; government; comment; economic; billion; new; bank; budget; debt; growth; jobs; market; million; spending

label     T1   T2   T3
cluster
0         70    2  143
1        129    3    3
2         15  201   34

              precision    recall  f1-score   support

          T1       0.33      0.33      0.33       214
          T2       0.02      0.01      0.02       206
          T3       0.14      0.19      0.16       180

    accuracy                           0.18       600
   macro avg       0.16      0.18      0.17       600
weighted avg       0.16      0.18      0.17       600
```

In [59]:
```
def cluster_kmean(train_data, test_data, num_clusters, min_df=1, stopw
    tfidf_vect = TfidfVectorizer(stop_words="english", min_df=1)
    dtm = tfidf_vect.fit_transform(train_data["text"])
    num_clusters = 3

    # initialize clustering model
    clusterer = KMeansClusterer(num_clusters, cosine_distance, repeats

    # samples are assigned to cluster labels
    # starting from 0
    clusters = clusterer.cluster(dtm.toarray(), assign_clusters=True)

    centroids = np.array(clusterer.means())

    # argsort sort the matrix in ascending order
    # and return locations of features before sorting
    # [:,::-1] reverse the order
    sorted_centroids = centroids.argsort()[:, ::-1]

    # The mapping between feature (word)
    # index and feature (word) can be obtained by
    # the vectorizer's function get_feature_names()
    voc_lookup = tfidf_vect.get_feature_names()

    for i in range(num_clusters):
        # get words with top 20 tf-idf weight in the centroid
        top_words = [voc_lookup[word_index] for word_index in sorted_c
        print("Cluster %d:\n %s " % (i, "; ".join(top_words)))

    test_dtm = tfidf_vect.transform(test_data["text"])

    predicted = [clusterer.classify(v) for v in test_dtm.toarray()]
    confusion_df = pd.DataFrame(list(zip(test_data["label"].values, pr

    print()
    # generate crosstab between clusters and true labels
    print(pd.crosstab(index=confusion_df.cluster, columns=confusion_df

    cluster_dict = {0:'T1', 1:"T2", 2:'T3'}

    # Map true label to cluster id
    predicted_target = [cluster_dict[i] for i in predicted]

    print()

    print(metrics.classification_report(test_data["label"], predicted_
```

In [60]:
```
# Clustering by Euclidean distance
cluster_kmean(train_data, test_data, num_clusters, min_df=1, stopwords
```

```
/Users/yashitavajpayee/opt/anaconda3/lib/python3.9/site-packages/skle
arn/utils/deprecation.py:87: FutureWarning: Function get_feature_name
s is deprecated; get_feature_names is deprecated in 1.0 and will be r
emoved in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)

Cluster 0:
 crash; bus; said; plane; police; train; car; accident; driver; road;
injured; says; people; killed; vehicle; man; passengers; hospital; cr
ashed; injuries
Cluster 1:
 said; oil; people; bp; rail; japan; cruise; water; new; airlines; mi
llion; spill; city; gulf; disaster; nuclear; pakistan; government; ea
rthquake; state
Cluster 2:
 percent; tax; said; year; rate; economy; obama; comment; economic; g
overnment; billion; bank; debt; growth; budget; new; market; spending
; jobs; unemployment

label     T1   T2  T3
cluster
0         62    0  74
1        138   18  99
2         14  188   7

              precision    recall  f1-score   support

          T1       0.46      0.29      0.35       214
          T2       0.07      0.09      0.08       206
          T3       0.03      0.04      0.04       180

    accuracy                           0.14       600
   macro avg       0.19      0.14      0.16       600
weighted avg       0.20      0.14      0.16       600
```

Q2: GMM Clustering (5 points)

Define a function `cluster_gmm(train_data, test_data, num_clusters, min_df = 10, stopwords = stopwords)` to redo Q1 using the Gaussian mixture model.

Requirements:

- To save time, you can specify the covariance type as `diag` .
- Be sure to run the clustering with different initiations to get stabel clustering results
- Your F1 score on the test set should be around `70%` or higher.

In [37]:
```
def cluster_gmm(train_data, test_data, num_clusters, min_df = 10, stop
    tfidf_vect = TfidfVectorizer(stop_words="english", min_df=10)
    dtm = tfidf_vect.fit_transform(train_data["text"])

    lowest_bic = np.infty
    best_gmm = None
    n_components_range = range(2, 5)
    cv_types = ['spherical', 'tied', 'diag']

    for cv_type in cv_types:
        for n_components in n_components_range:
            gmm = GaussianMixture(n_components=n_components, covarianc
            gmm.fit(dtm.toarray())
            bic = gmm.bic(dtm.toarray())
            if bic < lowest_bic:
                lowest_bic = bic
                best_gmm = gmm

    print('Best GMM is:', best_gmm)
    print()
    test_dtm = tfidf_vect.transform(test_data["text"])
    predicted = best_gmm.predict(test_dtm.toarray())

    new_df = pd.DataFrame(list(zip(test_data["label"].values, predicte
    merged_df = pd.crosstab(index=new_df.cluster, columns=new_df.actua
    print(merged_df)
    print()
    matrix = merged_df.idxmax(axis=1)
    final_predicted = [matrix[i] for i in predicted]
    classification_report = metrics.classification_report(test_data["l
    print(classification_report)
```

In [38]:
```
cluster_gmm(train_data, test_data, num_clusters, min_df = 10, stopworc
```

```
Best GMM is: GaussianMixture(covariance_type='diag', n_components=4)

actual_class   T1   T2   T3
cluster
0              70  160  108
1               1   43    2
2             138    3   70
3               5    0    0

              precision    recall  f1-score   support

          T1       0.66      0.67      0.67       214
          T2       0.53      0.99      0.69       206
          T3       0.00      0.00      0.00       180

    accuracy                           0.58       600
   macro avg       0.40      0.55      0.45       600
weighted avg       0.42      0.58      0.47       600


/Users/yashitavajpayee/opt/anaconda3/lib/python3.9/site-packages/skle
arn/metrics/_classification.py:1318: UndefinedMetricWarning: Precisio
n and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this beha
vior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/yashitavajpayee/opt/anaconda3/lib/python3.9/site-packages/skle
arn/metrics/_classification.py:1318: UndefinedMetricWarning: Precisio
n and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this beha
vior.
  _warn_prf(average, modifier, msg_start, len(result))
/Users/yashitavajpayee/opt/anaconda3/lib/python3.9/site-packages/skle
arn/metrics/_classification.py:1318: UndefinedMetricWarning: Precisio
n and F-score are ill-defined and being set to 0.0 in labels with no
predicted samples. Use `zero_division` parameter to control this beha
vior.
  _warn_prf(average, modifier, msg_start, len(result))
```

Q3: LDA Clustering (5 points)

Q3.1. Define a function `cluster_lda(train_data, test_data, num_clusters, min_df = 5, stopwords = stopwords)` to redo Q1 using the LDA model. Note, for LDA, you need to use `CountVectorizer` instead of `TfidfVectorizer` .

Requirements:

- Your F1 score on the test set should be around `80%` or higher
- Print out top-10 words in each topic
- Return the topic mixture per document matrix for the test set(denoted as `doc_topics`) and the trained LDA model.

Q3.2. Find similar documents

- Define a function `find_similar_doc(doc_id, doc_topics)` to find `top 3 documents` that are the most thematically similar to the document with `doc_id` using the `doc_topics` . (1 point)
- Return the IDs of these similar documents.
- Print the text of these documents to check if their thematic similarity.

Analysis:

You already learned how to find similar documents by using TFIDF weights. Can you comment on the difference between the approach you just implemented with the one by TFID weights?

In [43]:
```
def cluster_lda(train_data, test_data, num_clusters, min_df=5, stop_word
    if stop_words_list is None:
        stop_words_list = nltk.corpus.stopwords.words('english')
    stop = list(stop_words_list) + ['said']
    count_vect = CountVectorizer(min_df=5, max_df=0.95, stop_words=stop)
    dtm_train = count_vect.fit_transform(train_data["text"])
    dtm_test = count_vect.transform(test_data["text"])
    tf_feature_names = count_vect.get_feature_names()

    lda_model = LatentDirichletAllocation(n_components=num_clusters, max
    doc_topics = lda_model.transform(dtm_test)
    topic_final = doc_topics.argmax(axis=1)
    data_df = pd.DataFrame({"actual_class": test_data["label"], "cluster
    data_df_converted = pd.crosstab(index=data_df.cluster, columns=data_
    matrix = data_df_converted.idxmax(axis=1)
    final_predicted = [matrix[i] for i in topic_final]

    num_top_words = 10
    for topic_idx, topic in enumerate(lda_model.components_):
        print ("Topic %d:" % (topic_idx))
        words = [(tf_feature_names[i], topic[i]) for i in topic.argsort(
        print(words)
        print("\n")

    print(data_df_converted)
    print(metrics.classification_report(test_data["label"], final_predic

    return doc_topics, lda_model
```

In [44]:
```
num_clusters = 3
cluster_lda(train_data, test_data, num_clusters, min_df=5, stop_words_
```

```
/Users/yashitavajpayee/opt/anaconda3/lib/python3.9/site-packages/skle
arn/utils/deprecation.py:87: FutureWarning: Function get_feature_name
s is deprecated; get_feature_names is deprecated in 1.0 and will be r
emoved in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)

iteration: 1 of max_iter: 30, perplexity: 4255.7175
iteration: 2 of max_iter: 30, perplexity: 3739.2979
iteration: 3 of max_iter: 30, perplexity: 3591.4628
iteration: 4 of max_iter: 30, perplexity: 3540.3435
iteration: 5 of max_iter: 30, perplexity: 3517.2600
iteration: 6 of max_iter: 30, perplexity: 3505.7244
iteration: 7 of max_iter: 30, perplexity: 3499.2432
iteration: 8 of max_iter: 30, perplexity: 3494.9215
iteration: 9 of max_iter: 30, perplexity: 3491.7938
iteration: 10 of max_iter: 30, perplexity: 3489.1921
```

```
iteration: 11 of max_iter: 30, perplexity: 3487.0082
iteration: 12 of max_iter: 30, perplexity: 3485.1597
iteration: 13 of max_iter: 30, perplexity: 3483.4885
iteration: 14 of max_iter: 30, perplexity: 3481.8706
iteration: 15 of max_iter: 30, perplexity: 3480.2408
iteration: 16 of max_iter: 30, perplexity: 3478.5470
iteration: 17 of max_iter: 30, perplexity: 3476.7344
iteration: 18 of max_iter: 30, perplexity: 3474.8876
iteration: 19 of max_iter: 30, perplexity: 3473.1249
iteration: 20 of max_iter: 30, perplexity: 3471.4574
iteration: 21 of max_iter: 30, perplexity: 3469.8414
iteration: 22 of max_iter: 30, perplexity: 3468.2652
iteration: 23 of max_iter: 30, perplexity: 3466.8155
iteration: 24 of max_iter: 30, perplexity: 3465.6752
iteration: 25 of max_iter: 30, perplexity: 3464.8772
iteration: 26 of max_iter: 30, perplexity: 3464.2793
iteration: 27 of max_iter: 30, perplexity: 3463.8077
iteration: 28 of max_iter: 30, perplexity: 3463.4492
iteration: 29 of max_iter: 30, perplexity: 3463.1914
iteration: 30 of max_iter: 30, perplexity: 3463.0047
Topic 0:
[('new', 1714.2302146653092), ('com', 1401.9917030573843), ('rail', 1
151.2603232176557), ('service', 999.1362446393622), ('high', 949.2761
862671063), ('travel', 939.338326373009), ('one', 931.4673241267501),
('www', 877.6315665970654), ('cruise', 864.3272843396214), ('passenge
rs', 858.5578035556234)]


Topic 1:
[('people', 3352.634239772012), ('comment', 2976.9614799456513), ('oi
l', 1945.759418991494), ('news', 1720.0107573492187), ('one', 1529.89
35592809885), ('sign', 1504.4334921782422), ('please', 1382.090403182
392), ('users', 1342.1296752708347), ('water', 1336.1842289895558), (
'japan', 1322.7458640891764)]


Topic 2:
[('percent', 3967.1904562157297), ('year', 3160.2764519141424), ('tax
', 3019.5298172474145), ('would', 2706.005742910337), ('government',
2292.047486812791), ('obama', 2154.791179796024), ('economy', 2034.81
15277396666), ('new', 1827.7789156527053), ('billion', 1822.177349301
324), ('economic', 1804.5833127044218)]


actual_class   T1   T2   T3
cluster
0               27   18  126
1              174    3   42
2               13  185   12
              precision    recall  f1-score   support
```

```
              T1       0.79      0.81      0.80       214
              T2       0.88      0.90      0.89       206
              T3       0.74      0.70      0.72       180

        accuracy                           0.81       600
       macro avg       0.80      0.80      0.80       600
    weighted avg       0.81      0.81      0.81       600
```

Out[44]:
```
(array([[0.19921381, 0.2089335 , 0.59185269],
        [0.00346432, 0.21590973, 0.78062595],
        [0.0257079 , 0.0857817 , 0.8885104 ],
        ...,
        [0.13865422, 0.82685619, 0.03448959],
        [0.00605166, 0.98835983, 0.00558852],
        [0.04847507, 0.05780492, 0.89372002]]),
 LatentDirichletAllocation(evaluate_every=1, max_iter=30, n_components=3,
                           n_jobs=1, random_state=0, verbose=1))
```

In [69]:
```
f find_similar(doc_id, doc_topics):
  # Get the topic mixture of the target document
  target_doc_topic_mix = doc_topics[doc_id]

  # Calculate the cosine similarity of the target document's topic mixt
  similarity_scores = []
  for i, doc_topic_mix in enumerate(doc_topics):
      similarity_score = np.dot(target_doc_topic_mix, doc_topic_mix) /
      similarity_scores.append((i, similarity_score))

  # Sort the similarity scores in descending order
  similarity_scores = sorted(similarity_scores, key=lambda x: x[1], rev

  # Return the top 3 most similar document IDs
  docs = [score[0] for score in similarity_scores[1:4]]

  return docs
```

In [70]:
```
```

```
doc_topics[10:15]

doc_id = 11
idx = find_similar(doc_id, doc_topics)

print(test_data.text.iloc[doc_id])
print("Similar documents: \n")
for i in idx:
    print(i, test_data.iloc[i].text)
```

Out[70]:
```
array([[0.29767864, 0.24520652, 0.45711484],
       [0.00495174, 0.41766306, 0.5773852 ],
       [0.62835741, 0.00083028, 0.3708123 ],
       [0.07296908, 0.77729441, 0.14973651],
       [0.8245571 , 0.00147165, 0.17397125]])
```

```
obama says he s finding out whose ass to kick over gulf disasterbylin
e time tue jun am et is president obama bowing to criticism that he h
asn t shown enough emotion and outrage about the gulf of mexico oil s
pill in an interview with the today show s matt lauer this morning th
e president offered his most candid response yet about the disaster b
luntly telling lauer he s been talking to experts about whose ass to
kick when it comes to responsibility for the mess i was down there a
month ago before most of these talking heads were even paying attenti
on to the gulf a month ago i was meeting with fishermen down there st
anding in the rain talking about what a potential crisis this could b
e obama said defending his administration s handling of the spill and
i don t sit around just talking to experts because this is a college
seminar we talk to these folks because they potentially have the best
answers so i know whose ass to kick that s a pretty sharp response fo
r a president known for his cool headed approach to situations in rec
ent weeks as obama was assailed by critics for not being expressive e
nough in his response to the spill white house officials defended his
reaction by suggesting voters would prefer to see concrete actions ov
er empty method acting yet administration officials are not ignorant
of polls showing the nation less than thrilled with obama s handling
of the gulf according to the latest abc washington post poll more tha
n two thirds of those polled percent disapprove of the federal govern
ment s handling of the spill that s higher than the outrage over the
bush administration s handling of hurricane katrina holly bailey is a
senior political writer for yahoo news

Similar documents:

474 feds bp agrees to expedite oil spill payments washington the obam
a administration says bp has agreed to expedite the payment of claims
to businesses and individuals whose livelihoods have been disrupted b
y the gulf of mexico oil spill tracy wareing wehr ing who is with the
national incident command office told reporters in washington that th
e understanding on payment of claims came in a meeting wednesday with
```

bp executives including ceo tony hayward wareing said administration officials raised a pressing concern about the time bp has been taking to provide relief payments particularly to businesses in the stricken area she said the company will change the way it processes such claim s and will expedite payments among other things it will drop the curr ent practice of waiting to make such payments until businesses have c losed their books for each month

169 feds bp agrees to expedite oil spill payments the obama administr ation says bp has agreed to expedite the payment of claims to busines ses and individuals whose livelihoods have been disrupted by the gulf of mexico oil spill tracy wareing wehr ing who is with the national i ncident command office told reporters in washington that the understa nding on payment of claims came in a meeting wednesday with bp execut ives including ceo tony hayward wareing said administration officials raised a pressing concern about the time bp has been taking to provid e relief payments particularly to businesses in the stricken area she said the company will change the way it processes such claims and wil l expedite payments among other things it will drop the current pract ice of waiting to make such payments until businesses have closed the ir books for each month

248 ireland set for majority stakes in top banks home u s business wo rld entertainment sports tech politics science health opinion most po pular business video u s economy stock markets earnings opinion perso nal finance press releases taxes marketplace newsmakers search menu s earch type choose a search type from the items below all news yahoo n ews only news photos video audio news search trending now wikileaks m ichael vick amare stoudemire bob feller justin bieber los angeles lak ers mexico bellagio regis and kelly paul pierce sal alosi minnesota v ikings julian assange russia mariah carey cliff lee immigration obama nicole richie rip torn weekend edition the latest health leisure trav el and fitness news ireland set for majority stakes in top banks buzz up dji gspc ixic by steve slater steve slater wed nov pm et dublin re uters ireland is set to take a majority stake in top lender bank of i reland bkir i as part of a massive international bailout that could l eave the state with effective control of the country s top three bank s the state s ownership of bank of ireland could rise to near percent from percent now under the bailout put at up to billion euros billion and allied irish banks albk i could join anglo irish bank in being fu lly nationalized the european union and international monetary fund i mf have agreed to provide external assistance to ireland to shore up its banks and give them access to cheaper state funding billions of t he bailout could be used immediately to recapitalize the banks but mo st will be a backstop in case they need more in the future and to eas e funding strains irish officials have said they wanted to overcapita lize banks lenders are likely to be told to hold a core tier capital ratio of about percent giving a bigger cushion than most internationa l rivals to withstand future shocks and will need to draw on the back stop capital when the ratio drops below percent a person familiar wit

h the matter said bank of ireland could need over billion euros and a
ib even more and the government may be the only provider of funds in
the current environment i do not think they stand any chance of getti
ng all that privately said ciaran callaghan analyst at ncb in dublin
estimating that would leave the state with about an percent holding i
n bank of ireland lifting the core tier ratio of bank of ireland aib
and anglo irish bank would cost almost billion euros reuters calculat
ions showed including about billion for aib billion for bank of irela
nd and million for anglo irish bank more would be needed if preferenc
e shares and other types of capital were excluded but banks could be
given longer to reach that minimum capital level and be allowed to us
e retained earnings or tap outside investors to limit dilution a plun
ge in share prices this week has increased the dilution for sharehold
ers and the size of the government stake bank of ireland shares fell
percent to cents cutting its market value to under billion euros aib
shares lost over percent at one stage but ended percent higher valuin
g it at less than million euros both have lost over a third of their
value this week the government could be left with over percent of aib
and may only leave the shares listed to make it easier to sell down t
he stake in future analysts said irish prime minister brian cowen sai
d detail of the recapitalization had not been finalized extra cash co
uld be pumped into ireland s ailing banks as soon as this weekend the
irish independent reported on wednesday as well as providing capital
the government wants to shrink banks loan books to ease a funding str
ain which has intensified in the past six months after an exodus of d
eposits adding to lenders dependence on ecb funding which has risen t
o billion euros ireland could also impose a levy on banks as a term o
f the bailout and to ease a deadlock over the country s low corporati
on tax the irish times reported banks have been told to sell assets t
o focus on aiding the domestic economic recovery that could see aib b
eing told to restart a sale of its british business after halting the
process when failing to find a buyer some eu politicians are calling
for irish bank senior bondholders to share taxpayers pain and take ha
ircuts on the debt but most of that debt is explicitly guaranteed by
the irish government and bankers said such a move would damage prospe
cts for long term funding for ireland and other euro zone countries t
he banking crisis has rocked ireland and stretched its finances the d
eeply unpopular government on wednesday unveiled a billion euro four
year austerity plan of deep spending cuts to meet the terms of the ba
ilout additional reporting by carmel crimmins and alex chambers editi
ng by dan lalor and erica billingham euro buzz up explore related con
tent photo in slideshow bailed out bank of ireland flogs off art coll
ection photo icicles hang from a sculpture opposite the ulster bank o
n dame photo cranes stand above the unfinished anglo irish bank that
was to be allied irish banks no bonuses to allied irish banks says it
won t award bonuses to its employees following a full story ap photo
a woman walks past a branch of the bank of ireland in central photo b
ank of ireland to seek new capital injection anglo irish bank credit
rating is anglo irish bank corp had its long term counterparty credit
rating lowered to full story bloomberg irish government blocks aib bo

nuses london marketwatch allied irish banks said late monday that it has full story market watch aib credit institutions bill dublin ireland marketwire allied irish banks p l c aib full story marketwire photo removes reference to the irish constitution bank of ireland photo in this photo taken on nov an exit sign bank of ireland treasuries gain amid irish bank s treasuries gained as standard poor s ratings services downgraded anglo full story bloomberg more on europe eu leaders set crisis fund ecb boosts capital reuters europe throws euro fresh lifeline afp russia in contact with shuttle after glitch reuters more business video talking numbers cnbc business video oil gold and silver tomorrow cnbc business video morgan stanley what s next cnbc comments show newest firstoldest firsthighest ratedmost replied sort post a comment comments of prevnextlast users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment alexander sun nov pm pst report abuse a wise but late move users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment jim sun nov pm pst report abuse the irish central banks wants easy money just like the bernanke fed they want a ponzi scheme just like the us fed they want the irish taxpayer to bail out their banks and bondholders what a bunch of crooks we have in govt these days users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment leonardo fri nov am pst report abuse ah the luck a the irish they will be lucky if they can trade their bonds for potatos users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment what fri nov am pst report abuse maybe cause they are always drunk you know the mic s users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment fri nov am pst report abuse a couple of years ago the idiot conservative talk show hosts were all saying we needed to model our tax system after the irish they said that we need to lower our corporate taxes to the level of ireland to encourage businesses to move to the us clueless as usual users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment lisa fri nov am pst report abuse halt inflation temp of the billion disburse part of the money to the citizens of the united states from the ceo to the street sweaper all an equal share make the people responsiable very easy to do the problem is cash flow the people who have it only very few earned it time to move the econmy users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment bob ross fri nov am pst report abuse don t make the mistake we did kill the bankers users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment karl krupp fri nov am pst report abuse them dirty leprechauns stole it while me was drunk users liked this comment please sign in to rate this comment up please sign in to rate this comment down users disliked this comment localdiner fri nov am pst report abuse of course the bond holders should get

a haircut a bond is a loan and the bond holders lent money to a shaky business whether they knew it was shaky or not is immaterial you make a loan and with it comes the risk that it might go bad that is why th ere are interest rate premiums too bad for the bondholders but they m ade the loan and took the risk better they should get hit since they were the ones that hoped to profit rather than some poor taxpayer who is not party to the transaction capitalism only works if there is the possibility of failure to go with the possibility of profit otherwise it is just an old boy crony system users liked this comment please si gn in to rate this comment up please sign in to rate this comment dow n users disliked this comment enforce this fri nov am pst report abus e how dare ap use the word spook in any article im calling the aclu c omments of prevnextlast post a comment sign in to post a comment or s ign up for a free account most viewed business after dozens of deaths drop side cribs outlawed ap paris taps firm to run new electric car s hare plan ap irs audits jump by percent wealthiest targeted ap fyi on line retailers offer free shipping friday ap senate gop leader offers two month spending bill ap all most viewed daily features all comics opinions editorials diverse views on news from the right left and cen ter all opinion elsewhere on the web time com congress tells commerci als to quiet down time com google s nexus s the best android smart ph one right now foxbusiness tax cuts in jeopardy top stories final hurd le for tax bill is the end in sight despite bloodshed obama touts afg han war progress wikileaks assange walks free on bail in london is pr ivatizing government and maybe even the moon a way to trim deficit ho nda to recall million fit subcompacts globally senators vote to ban e armarks â then grab them irs audits jump by percent wealthiest target ed photos freezing temps turn lighthouse into ice sculpture cnn host larry king hands over his vintage microphone fda says avastin does no t work for breast cancer denied a bank account instant approval no cr edit checks visa debit card online bill pay www readydebit com bank a ccount affordable loan modify do you qualify find instantly million h omeowners can benefit save my home org free banking resource comprehe nsive objective and free banking resource online www bankrate com fea tured watch tackling debttips for paying down your debt one step at a time protect your moneyfour ways to avoid being madoffed social secur ity necessary commentary cheering the attack on social security free shipping fridayhow to make the most of free shipping day educationyah oo your recession weaponsee how education can help you avoid the nega tive effects of the recession is an mba for you what you need to know before you choose your online mba program yahoo dow nasdaq s p yr bon d oil gold fed proposes cent cap on merchant debit fees research in m otion profit jumps percent oracle reports profit jump bucking industr y fears yahoo news navigation home u s business world entertainment s ports tech politics science health travel most popular odd news opini on yahoo news network rss news alerts weather alerts site map help fe edback errors or delays in the content or for any actions taken in reliance

```
thereon
```

Q4 (Bonus): Find the most significant topics in a document

A small portion of documents in our dataset have multiple topics. For instace, consider the following document which has topic T2 and T3. The LDA model returns two significant topics with probabilities 0.355 and 0.644. Can you describe a way to find out most significant topics in documents but ignore the insignificant ones? In this example, you should ignore the first topic but keep the last two.

- Implement your ideas
- Test your ideas with the test set
- Recalculate the precision/recall/f1 score for each label.

In [135]:
```
(test_data.reset_index()).iloc[12:13]
doc_topics[12]
```

Out[135]:

	index	text	T1	T2	T3
12	12	white house to dole out billion for fast train...	0	1	1

Out[135]:
```
array([0.00091134, 0.35500994, 0.64407872])
```

In [*]:
```
if __name__ == "__main__":

    # Due to randomness, you won't get the exact result
    # as shown here, but your result should be close
    # if you tune the parameters carefully
    train = pd.read_csv("hw5_train.csv")
    train.head()

    test = pd.read_csv("hw5_test.csv")
    test.head()


    # Q1
    cluster_kmean(train_data, test_data, num_clusters, min_df=5, metri

    # Q2

    cluster_gmm(train_data, test_data, num_clusters, min_df = 10, stop
```

Out[76]:

	text
0	blm in wyo to begin deciding on backlogged lea...
1	report amtrak loss comes to per passenger u s ...
2	medicare key in races washington an upset vict...
3	sunnyvale bicyclist dies of injuries suffered ...
4	mozambique upbeat on debt crisis investors not...

Out[76]:

	text	T1	T2	T3
0	child asylum seekers targeted in home office b...	0	1	0
1	obama acknowledges economic stress not so long...	0	1	0
2	help not soonbyline by ted ralltime fri jul pm...	0	1	0
3	un pakistan flood misery exceeds tsunami haiti...	1	0	0
4	new home sales plunge new home sales plunge we...	0	1	0

In [ ]: