Arabic Text Classification - Moroccan News

In [673]:
```
!python __init__.py
```

In [674]:
```
from __init__ import *
```

In [675]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import warnings
import pre_processing as pp  # local module
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import cross_val_score
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
from lxml import html
import requests
import nltk
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
import matplotlib.pylab as plt
from gensim.models import word2vec
from sklearn.manifold import TSNE
from sklearn.metrics import confusion_matrix
from sklearn.feature_extraction.text import CountVectorizer
```

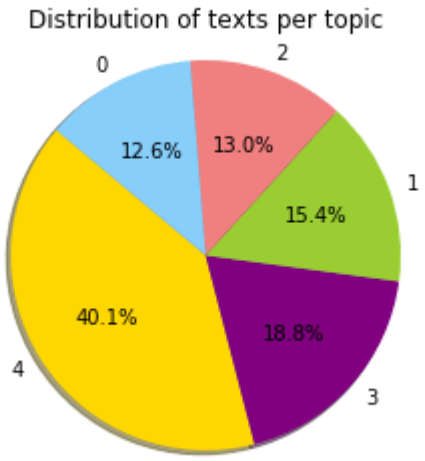
In [676]:
```
data=pd.read_csv(r'arabic_classification_train.csv')
```

In [677]:
```
data=data.dropna()
```

Visualization

In [708]:
```
topics=data.groupby('label',as_index=False)['text'].count().sort_values(['tex
t'],ascending = False)
```

In [716]:
```
colors = ['gold','purple', 'yellowgreen', 'lightcoral', 'lightskyblue']
sizes=topics['text']
labels=topics['label']
# Plot
plt.pie(sizes,  labels=labels, colors=colors,
autopct='%1.1f%%', shadow=True, startangle=140)
plt.title('Distribution of texts per topic')
plt.axis('equal')
plt.show()
```



Preprocessing

In [680]:
```
# wpt = nltk.WordPunctTokenizer()
```

In [681]:
```
def tokenize_text(text):
    tokens = nltk.word_tokenize(text)
    tokens = [token.strip() for token in tokens]
    return tokens
```

In [682]:
```
# tokenized_sent=data['text'].apply(lambda x: wpt.tokenize(x))
```

In [683]:
```
page = requests.get('https://ar.wikipedia.org/wiki/%D8%A7%D8%B3%D8%AA%D8%A8%D8%B9%D8%A7%D8%AF_%D8%A7%D9%84%D9%83%D9%84%D9%85%D8%A7%D8%AA_%D8%A7%D9%84%D8%B4%D8%A7%D8%A6%D8%B9%D8%A9')
tree = html.fromstring(page.content)
```

In [684]:
```
stopword_list = list(tree.xpath('//li/text()'))
```

In [685]:
```
stopword_list.append('أن')
stopword_list.append('انه')
stopword_list.append('مغرب')
```

In [686]:
```
'انه' in stopword_list
```

Out[686]: True

In [687]:
```
def remove_stopwords(text):
    tokens = tokenize_text(text)
    filtered_tokens = [token for token in tokens if token not in stopword_list
]
    filtered_text = ' '.join(filtered_tokens)
    return filtered_text
```

In [688]:
```
def remove_taarif(text):
    tok_sent=tokenize_text(text)
    return(' '.join([re.sub('^ال','',tok) for tok in tok_sent ]))
```

In [689]:
```
# data['text']=data['text'].apply(lambda x: remove_stopwords(x))
```

In [690]:
```
data['text'].head()
```

Out[690]:
```
0    يخوض المنتخب الوطني داخل القاعة مباراتين إعداد...
1    تمكنت مصالح الدرك الملكي لطنجة والعرائش من إيق...
2    بعد ما أعلنت نقابة الاتحاد الوطني للشغل بالمغر...
3    أخبارنا المغربية سناء الوردي عرت صحيفة القدس ا...
4    تمكنت المصالح الجمركية بعد عصر اليوم الجمعة بم...
Name: text, dtype: object
```

In [691]:
```
ArListem=ArabicLightStemmer();
```

In [692]:
```
def preprocessing_arabic(text):
    text=remove_taarif(text)
    ArListem.lightStem(text)
    text=ArListem.get_normalized()
    text=ArListem.get_stem()
    text=strip_tashkeel(text)
    text=strip_tatweel(text)
    text=normalize_lamalef(text)
    text=normalize_hamza(text)
    text=normalize_spellerrors(text)
    return(text)
```

In [693]:
```
data['text']=data['text'].apply(lambda x: remove_stopwords(x))
```

In [694]:
```
data['text']=data['text'].apply(lambda x:preprocessing_arabic(x))
```

In [695]:
```
# data['text']=data['text'].apply(lambda x: remove_stopwords(x))
```

In [696]:
```
data['text'][5]
```

Out[696]:
```
'نسبه للجزاءريين مقيمين بالمغرب والمغاربه مقيمين بالجزاءر تعتبر مساله حدود مغلقه قضيه سياسيه وليس للشعوب يد و
داءما طرح موضوع شارع اي نقاش علاقات ثناءيه بين بلدين جارين يترتب عنه حديث تعثر مشروع اتحاد مغرب عربي
بسبب توترات سياسيه اشراق طوافي طالبه مغربيه لديها اصدقاء جاليه جزاءريه بالدار بيضاء تقول شعوب يمكنها تميز بي
ن علاقات جوار والقضايا سياسيه وتضيف طوافي شخصيا قررت توقف مناقشه سياسه جزاءريين اعرفهم انها مساله شديده
تعقيد وعوض اثاره عداوه اصبحت مساله حدود مصدرا للفكاهه ديوانه بطوله كوميدي مغربي حسن فد والجزاءري عبد قا
در سيكتور دوري جمركيين رض خلال شهر رمضان قناه مغربيه ثانيه سلسله تطرقت لاغلاق حدود يتواصل لقرابه سنه
قالب فكاهي ومع فالجدل سياسي بين حكومتين ليس تاثير ذكر علاقات شخصيه بين شعبين مخاوف بان رعايا بلدين يعيشو
ن بلد اخر يعيشون عزله او تمييزا اساس صحه سياق تقول طوافي تحضر شهاده دكتوراه داءما اجد بان جزاءريين يعيش
ن بين ظهرانينا يشعرون بانهم جنسيه مختلفه حسين بولمان مقاول مغربي شاب لديه مصالح تجاريه جزاءر نفس شعور ل
مغاربيه داءما يرحب بي جزاءريين اشعر ابدا باي تمييز سواء عموم ناس او سلطات كلما ذهبت ويقول لما قمت باول ر
حله لي اشعر بالغربه انني سافرت بلد بلد وحتي طاءره يقول بولمان انها تقلع دار بيضاء مغرب وتحط بالدار بيضاء ج
زاءر اما جزاءري سمير عبد مالك عامل قطاع صحافه بالمغرب انه تعترضه ايه مشاكل للاندماج عالم اعلامي عبر حدو
د انه يجد سهوله انتقال مهنه مهنه لتطوير مساره مهني نفس انطباع يبديه مغاربه مقيمون بالجزاءر ندي شابه استقرت مء
خرا مدينه جزاءر زوجها قالت انها تعاني ابدا اي تمييز وتقول ندي جزاءريون طيبون ويهتمون بنا ويسالوننا ايضا ان كن
اندمجنا مجتمع وهكذا ينبغي تكون امور فنحن تقريبا نفس شعب ولدينا كثير امور مشتركه وتضيف خلافات سياسيه علاق
ه بنا مساله حدود مغلقه والازعاج تسببه للعاءلات مصدر قلق بالنسبه لياسمينه سي عبد رحمان عضو جمعيه مغرب كبير
وتقول وجده اشخاص لديهم اقرباء جزاءر تفصلهم سوي كيلومترات جهه اخري حدود وضع ليس مريحا شعبان يريدان اقت
راب بعضهما بعض ومضت تقول نكون اتصال بعضنا بعض نتفاهم بشكل جيد نعتقد اننا شعب نفس تقاليد ونفس ثقافه د
فء يتسم شارع بدا يسري ساحه دبلوماسيه فالامور اخذت منحي ايجابيا شهور اخيره فالعلاقات جزاءريه مغربيه بدات تتح
سن بشكل جيد يناير زار وزير خارجيه مغربي سعد دين عثماني جزاءر عاصمه وجاءت زياره اولي نوعها لوزير خارج
يه مغربي لاحياء اتحاد مغاربي اتحاد مغرب عربي بدوره صرح رءيس وزراء جزاءري احمد اويحي ليوميه خبر مارس
حدود بين جزاءر والمغرب ست فتح طال زمن او قصر امور تسير يوم بشكل ممتاز اساس زيارات متبادله واللقاءات جا
نبين يشير دفء علاقات لهذا فان فتح حدود بات ضروريا خاصه واننا بلدين جارين مطلوب منا تفاهم معا بالنظر امور
وحدنا كاللغه والدين والتقاليد مشتركه وتاريخنا ومستقبلنا مشترك اواخر شهر ماضي جدد ملك محمد سادس نداءه دفع بالات
حاد مغاربي وضع جمود حالي وضع ديناميه شانه يساعدنا تحقيق تنميه مستدامه ومدمجه ملك خطاب تلفزيوني بمناسبه عي
د عرش مغرب سيواصل مساعيه لتعزيز علاقات ثناءيه كافه شركاءه مغاربيين بما فيهم جارتنا وشقيقتنا جزاءر اجل استجا
به تطلعات ملحه والمشروعه لشعوب منطقه شهدت تجاره بين جزاءر وباقي بلدان اتحاد مغرب عربي تحسنا بمعدل ماءه
ويبقي مغرب شريك تجاري اول للجزاءر سنه ماضيه تقريرا جديدا صدر مطلع شهر جاري بنك افريقي للتنميه يشير منط
قه مغاربيه جوار اقل اندماجا اقتصاديا عالم واشار بنك افريقي حدود مغلقه بين جزاءر والمغرب اكبر عاءق امام اتحاد م
غرب عربي جامد لسنوات وبالرغم تعثر سياسي فان امكانيات هامه بحسب اقتصادي بنك افريقي للتنميه امانويل سانتي وي
نيف سانتي ظرفيه سياسيه جديده دول شمال افريقيا والازمه اوروبا ترغم بلدان تنويع اسواق توفر فرصه ذهبيه لاعاده ت
ركيز اجنده اندماج اقليمي كمحرك للنمو بالنسبه لكافه بلدان ارشيف عشرات نشطاء مجتمع مدني اجتمعوا بنواكشوط ديسمب
ر ماضي مءتمر اتحاد شباب اورو مغاربي ارشيف عشرات نشطاء مجتمع مدني اجتمعوا بنواكشوط ديسمبر ماضي مءتمر
اتحاد شباب اورو مغاربي ومهما حدث بين حكومتي مغرب والجزاءر فان مواطني بلدين شقيقين لهم نفس هدف حسب تغ
زوت غزالي حركه جزاءريه للشباب مستقل اجل تغيير والعضو مءسس اتحاد شباب اورو مغاربي تريده شعوب فتح حدود
'ليس جميلا تري اناسا يعيشون كيلومترات بعضهم بعض قادرين عبور حدود بسبب صراع سياسي بين بلدين متابعه
```

TFIDF

In [697]:
```
tf_idf = TfidfVectorizer(min_df=0.1, max_df=0.7, norm='l2', use_idf=True, smoo
th_idf=True)
```

In [698]:
```
# cv = CountVectorizer(ngram_range=(1,1))
# cv1 = cv.fit_transform(data['text'].head(10000))
# cv1 = cv1.toarray()

# vocab = cv.get_feature_names()
# dfTF = pd.DataFrame(cv1, columns=vocab)

# dfTF



#cv = CountVectorizer(ngram_range=(1,1))
cv1 = tf_idf.fit_transform(data['text'].head(10000))
cv_unsupervised=cv1
cv1 = cv1.toarray()

vocab = tf_idf.get_feature_names()
dfTF = pd.DataFrame(cv1, columns=vocab)

# dfTF
```

In [699]:
```
freq=[]
for col in dfTF.columns:
    freq.append(sum(dfTF[col]))
tot_freq=zip(dfTF.columns,freq)
```

In [700]:
```
tot_freq=pd.DataFrame(tot_freq, columns=['col','freq'])
tot_freq.sort_values(by='freq', ascending=False)
```

Out[700]:

	col	freq
79	فريق	712.555046
105	مغرب	663.464441
107	مغربيه	642.856023
50	خلال	597.809633
18	انه	575.092150
93	مباراه	553.668830
32	بين	553.496553
69	عام	496.903725
62	سنه	482.658144
47	حكومه	465.101387
106	مغربي	456.932282
57	رءيس	455.673354
2	اجل	455.215451
124	وطني	452.765044
98	محمد	442.874852
112	منتخب	430.192543
127	يوم	423.707739
103	مصادر	414.214878
20	او	413.228979
17	ان	407.276445
7	اذ	402.179316
71	عبد	392.963502
91	ماضي	390.256071
14	امام	386.361894
59	رياضي	380.267079
97	مجموعه	371.209904
81	قدم	367.685015
0	اتحاد	366.233853
70	عامه	364.289776
125	وطنيه	341.735362
...	...	...
95	مجال	212.123500
82	قرار	209.776377
119	والتي	209.653749
66	ضمن	208.306811

	col	freq
86	كره	206.917008
9	اسبوع	203.949383
61	سبت	203.449249
58	رغم	201.374874
67	طرف	199.856038
25	بان	198.230687
44	جهه	197.642448
42	جمعه	197.405832
46	حسب	196.897090
1	اثنين	196.151479
37	ثلاثه	191.917021
8	اربعاء	191.735153
55	دين	188.604003
51	خميس	188.510432
75	عديد	188.342992
123	وضع	187.266619
74	عدم	186.465858
39	جانب	184.592329
24	ايضا	184.030758
26	بدايه	183.621501
85	كبيره	178.461840
120	وان	173.517036
109	مقبله	172.012213
10	اضافه	167.210343
121	واوضح	166.891559
118	واكد	150.764158

128 rows × 2 columns

In [701]:
```
tot_freq.sort_values(by='freq', ascending=False).head(50)
```

Out[701]:

	col	freq
79	فريق	712.555046
105	مغرب	663.464441
107	مغربيه	642.856023
50	خلال	597.809633
18	انه	575.092150
93	مباراه	553.668830
32	بين	553.496553
69	عام	496.903725
62	سنه	482.658144
47	حكومه	465.101387
106	مغربي	456.932282
57	رءيس	455.673354
2	اجل	455.215451
124	وطني	452.765044
98	محمد	442.874852
112	منتخب	430.192543
127	يوم	423.707739
103	مصادر	414.214878
20	او	413.228979
17	ان	407.276445
7	اذ	402.179316
71	عبد	392.963502
91	ماضي	390.256071
14	امام	386.361894
59	رياضي	380.267079
97	مجموعه	371.209904
81	قدم	367.685015
0	اتحاد	366.233853
70	عامه	364.289776
125	وطنيه	341.735362
104	مصدر	341.524336
3	احد	326.447746
87	لاعب	321.268017
16	امس	315.538947
48	خاصه	314.296945

	col	freq
65	صباح	312.141157
108	مقبل	310.981409
21	اول	309.496553
5	اخري	306.737710
96	مجلس	306.504123
22	اولي	303.013504
40	جديد	299.562688
100	مدينه	299.298204
122	وذلك	297.069186
73	عدد	296.591371
49	خبارنا	295.751838
88	لاعبين	293.127198
23	اي	291.223967
113	موسم	290.849483
68	عالم	285.977517

In [703]:
```
similarity_matrix = cosine_similarity(dfTF.head(500))
similarity_df = pd.DataFrame(similarity_matrix)
```

In [704]:
```
similarity_df
```

Out[704]:

	0	1	2	3	4	5	6	7	8
0	1.000000	0.024856	0.090484	0.034854	0.078443	0.076093	0.399348	0.050340	0.082101
1	0.024856	1.000000	0.000000	0.052163	0.000000	0.107708	0.089732	0.371862	0.059236
2	0.090484	0.000000	1.000000	0.269956	0.074222	0.255201	0.079843	0.063682	0.248471
3	0.034854	0.052163	0.269956	1.000000	0.438587	0.427592	0.109922	0.054155	0.217367
4	0.078443	0.000000	0.074222	0.438587	1.000000	0.225973	0.055397	0.146411	0.134581
5	0.076093	0.107708	0.255201	0.427592	0.225973	1.000000	0.131874	0.059254	0.210494
6	0.399348	0.089732	0.079843	0.109922	0.055397	0.131874	1.000000	0.094777	0.219918
7	0.050340	0.371862	0.063682	0.054155	0.146411	0.059254	0.094777	1.000000	0.045264
8	0.082101	0.059236	0.248471	0.217367	0.134581	0.210494	0.219918	0.045264	1.000000
9	0.362184	0.072355	0.137831	0.099812	0.039027	0.033663	0.241396	0.037581	0.146759
10	0.261134	0.000000	0.154919	0.286517	0.339828	0.115144	0.147130	0.166092	0.165732
11	0.000000	0.000000	0.202398	0.315522	0.000000	0.090009	0.022115	0.024941	0.035679
12	0.054500	0.072073	0.085699	0.253210	0.082700	0.301397	0.190959	0.000000	0.067101
13	0.512349	0.021322	0.152204	0.067068	0.074333	0.153189	0.249498	0.014136	0.115322
14	0.073939	0.000000	0.292348	0.192100	0.053633	0.243937	0.226634	0.086679	0.191470
15	0.078866	0.185537	0.142203	0.116210	0.030009	0.237817	0.075730	0.293924	0.272106
16	0.155845	0.329440	0.068455	0.061923	0.189080	0.081275	0.119227	0.316232	0.113056
17	0.418233	0.000000	0.078220	0.165232	0.231999	0.223561	0.424952	0.018236	0.169210
18	0.051891	0.000000	0.106429	0.258771	0.147675	0.142096	0.058051	0.065471	0.042954
19	0.079521	0.188918	0.256224	0.278543	0.083877	0.359488	0.271444	0.056922	0.181963
20	0.374626	0.090578	0.241284	0.234380	0.218833	0.282163	0.298357	0.155304	0.323912
21	0.020551	0.000000	0.186170	0.156198	0.074714	0.129305	0.149051	0.000000	0.332733
22	0.156847	0.071074	0.236043	0.091873	0.073392	0.158942	0.141385	0.019714	0.282979
23	0.075270	0.053354	0.118631	0.120043	0.055094	0.075248	0.165316	0.072106	0.588408
24	0.167931	0.000000	0.116678	0.038322	0.015524	0.106922	0.285601	0.017739	0.155748
25	0.048524	0.000000	0.030102	0.084797	0.000000	0.195569	0.122486	0.000000	0.078568
26	0.016964	0.109250	0.021635	0.051170	0.000000	0.123821	0.103901	0.000000	0.050719
27	0.021211	0.093502	0.231140	0.352842	0.201422	0.341719	0.096185	0.028512	0.158040
28	0.525180	0.000000	0.172081	0.343706	0.208081	0.158703	0.419638	0.092021	0.221712
29	0.074077	0.040851	0.117406	0.279861	0.198902	0.206148	0.171502	0.009281	0.169510
...	...	...	...	...	...	...	...	...	...
470	0.039583	0.000000	0.126510	0.658642	0.232036	0.288850	0.082675	0.022535	0.455421
471	0.085813	0.043947	0.079905	0.247671	0.264876	0.224315	0.282812	0.029945	0.310385
472	0.040199	0.226075	0.052785	0.079857	0.036756	0.135745	0.196192	0.115986	0.129469
473	0.239030	0.000000	0.113895	0.181599	0.000000	0.123955	0.230036	0.054311	0.195443

	0	1	2	3	4	5	6	7	8
474	0.039430	0.000000	0.057265	0.422299	0.032216	0.202923	0.051330	0.025756	0.152670
475	0.115682	0.148982	0.049066	0.143737	0.028886	0.133996	0.327934	0.061549	0.233437
476	0.024825	0.000000	0.381805	0.140007	0.000000	0.145053	0.028469	0.000000	0.093047
477	0.081735	0.000000	0.178187	0.217455	0.018194	0.114738	0.313355	0.130270	0.248256
478	0.000000	0.336801	0.102675	0.127742	0.000000	0.233610	0.094156	0.446264	0.000000
479	0.080601	0.000000	0.172913	0.052218	0.077272	0.166856	0.261314	0.100528	0.374916
480	0.144806	0.000000	0.062414	0.077190	0.186951	0.207347	0.088930	0.109308	0.111415
481	0.064539	0.031337	0.105355	0.182104	0.093855	0.143329	0.037592	0.064036	0.283124
482	0.052866	0.000000	0.000000	0.027559	0.000000	0.028986	0.044357	0.000000	0.046506
483	0.817787	0.025765	0.084077	0.055476	0.056564	0.086678	0.356896	0.019046	0.059481
484	0.513194	0.090116	0.079419	0.111965	0.122095	0.127625	0.329840	0.026588	0.135863
485	0.131332	0.095550	0.032177	0.181867	0.090781	0.194916	0.467371	0.060186	0.362909
486	0.076294	0.000000	0.207982	0.275786	0.125505	0.211068	0.043125	0.085832	0.187108
487	0.560945	0.000000	0.116897	0.181420	0.171275	0.205051	0.355144	0.030825	0.115225
488	0.012449	0.258954	0.085980	0.160835	0.000000	0.201723	0.041181	0.029198	0.189555
489	0.044649	0.000000	0.117771	0.070256	0.109027	0.147071	0.241999	0.000000	0.229337
490	0.117217	0.000000	0.150462	0.238483	0.154241	0.199189	0.325055	0.134896	0.298627
491	0.070514	0.000000	0.192583	0.272885	0.089885	0.251713	0.000000	0.000000	0.048608
492	0.080346	0.066866	0.090622	0.132631	0.134506	0.157299	0.356564	0.059201	0.215244
493	0.106747	0.000000	0.083129	0.097092	0.000000	0.198344	0.156038	0.032768	0.136809
494	0.034963	0.000000	0.160257	0.491759	0.315915	0.476321	0.090293	0.043498	0.103972
495	0.000000	0.000000	0.174404	0.115564	0.062515	0.104397	0.023042	0.145883	0.283151
496	0.124626	0.000000	0.103966	0.106219	0.031490	0.114005	0.151343	0.017992	0.136669
497	0.045644	0.000000	0.036723	0.094447	0.254992	0.098027	0.156364	0.025515	0.089650
498	0.037084	0.000000	0.064639	0.283795	0.290343	0.152718	0.000000	0.035092	0.000000
499	0.074748	0.049799	0.037900	0.088745	0.115018	0.150869	0.220925	0.222312	0.191857

500 rows × 500 columns

Word Embedding

Robust Word2Vec Models with Gensim

In [706]:
```
wpt = nltk.WordPunctTokenizer()
tokenized_corpus = [wpt.tokenize(document) for document in data['text'].head(1
0000)]

# Set values for various parameters
feature_size = 100 # Word vector dimensionality
window_context = 30 # Context window size
min_word_count = 1 # Minimum word count
sample = 1e-3 # Downsample setting for frequent words
w2v_model = word2vec.Word2Vec(tokenized_corpus, size=feature_size,window=windo
w_context, min_count=min_word_count,sample=sample, iter=50)


# view similar words based on gensim's model
similar_words = {search_term: [item[0]
                   for item in w2v_model.wv.most_similar([search_term],topn=5
)]
                       for search_term in [ 'عربي' ,'ملك' ,'حكومه' ,'فريق' ,'محمد'
,'درهم' ,'كره','ه']}
similar_words
```

Out[706]:
```
{'محمد': ['سعديين', 'وينقلها', 'ضبطتا', 'بفتاتين', 'قاصرتين'],
 'فريق': ['للفريق', 'بالفريق', 'نادي', 'وداد', 'لاعبين'],
 'حكومه': ['للحكومه', 'حكومات', 'حكوميه', 'حكومته', 'بنكيران'],
 'ملك': ['جلالته', 'للتوجيهات', 'عاهل', 'صاحب', 'ساجلب'],
 'عربيه': ['عرب', 'خليج', 'سعوديه', 'ولشعوبهم', 'والاردن'],
 'كره': ['بكره', 'لكره', 'وكره', 'موطء', 'للمساحات'],
 'درهم': ['سنتيم', 'والجنايات', 'دولار', 'دراهم', 'وانكارهم']}
```

In [ ]:

In [ ]:
```
words = sum([[k] + v for k, v in similar_words.items()], [])
wvs = w2v_model.wv[words]
tsne = TSNE(n_components=2, random_state=0, n_iter=10000, perplexity=1)
np.set_printoptions(suppress=[True])
T = tsne.fit_transform(wvs)
labels = words
plt.figure(figsize=(10, 10))
plt.scatter(T[:, 0], T[:, 1],s=30, c='orange', edgecolors='r')
#plt.xlim(-400,600)
#plt.ylim(1000,1000)
for label, x, y in zip(labels, T[:, 0], T[:, 1]):
    plt.annotate(label, xy=(x+1, y+1), xytext=(1, 1), textcoords='offset pixe
ls')
plt.rc('font',size=10)
# plt.yscale('symlog')
# plt.xscale('symlog')
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

The Skip-Gram Model

Extract each unique word from our vocabulary and assign a unique identifier.

In [717]:
```
from keras.preprocessing import text
```

```
Using TensorFlow backend.
```

In [720]:
```
norm_corpus=data['text'].head(10000)
```

In [721]:
```
tokenizer = text.Tokenizer()
tokenizer.fit_on_texts(norm_corpus)
word2id = tokenizer.word_index
id2word = {v:k for k, v in word2id.items()}
vocab_size = len(word2id) + 1
embed_size = 100
wids = [[word2id[w] for w in text.text_to_word_sequence(doc)] for doc in norm_
corpus]
print('Vocabulary Size:', vocab_size)
print('Vocabulary Sample:', list(word2id.items())[:10])
```

```
Vocabulary Size: 125099
Vocabulary Sample: [('فريق', 1), ('مغرب', 2), ('خلال', 3), ('انه', 4), ('بين', 5),
('مباراه', 6), ('مغربيه', 7), ('عام', 8), ('وطني', 9), ('سنه', 10)]
```

Build a Skip-Gram [(target, context), relevancy] Generator

In [722]:
```
from keras.preprocessing.sequence import skipgrams
```

In [ ]:
```
# generate skip-grams
skip_grams = [skipgrams(wid, vocabulary_size=vocab_size, window_size=10)
for wid in wids]
# view sample skip-grams
pairs, labels = skip_grams[0][0], skip_grams[0][1]
for i in range(10):
    print("({:s} ({:d}), {:s} ({:d})) -> {:d}".format(id2word[pairs[i][0]], pa
irs[i][0],id2word[pairs[i][1]], pairs[i][1],labels[i]))
```

Build the Skip-Gram Model Architecture

In [ ]:
```
from keras.layers import Dot
from keras.layers.core import Dense, Reshape
from keras.layers.embeddings import Embedding
from keras.models import Sequential
from keras.models import Model
```

In [ ]:
```
# build skip-gram architecture
word_model = Sequential()
word_model.add(Embedding(vocab_size, embed_size, embeddings_initializer="gloro
t_uniform", input_length=1))
word_model.add(Reshape((embed_size, )))
context_model = Sequential()
context_model.add(Embedding(vocab_size, embed_size,embeddings_initializer="glo
rot_uniform",input_length=1))
context_model.add(Reshape((embed_size,)))
model_arch = Dot(axes=1)([word_model.output, context_model.output])
model_arch = Dense(1, kernel_initializer="glorot_uniform",activation="sigmoid"
)(model_arch)
model = Model([word_model.input,context_model.input], model_arch)
model.compile(loss="mean_squared_error", optimizer="rmsprop")

# view model summary
print(model.summary())

# visualize model structure
from IPython.display import SVG
from keras.utils.vis_utils import model_to_dot
SVG(model_to_dot(model, show_shapes=True, show_layer_names=False,rankdir='TB',
dpi=65).create(prog='dot', format='svg'))
```

Train the Model

In [ ]:
```
import time
ts = time.time()
for epoch in range(1, 10):
    loss = 0
    try:
        for i, elem in enumerate(skip_grams):
            pair_first_elem = np.array(list(zip(*elem[0]))[0], dtype='int32')
            pair_second_elem = np.array(list(zip(*elem[0]))[1], dtype='int32')
            labels = np.array(elem[1], dtype='int32')
            X = [pair_first_elem, pair_second_elem]
            Y = labels
            loss += model.train_on_batch(X,Y)
    except:
        pass
    print('Epoch:', epoch, 'Loss:', loss)
te = time.time() - ts
print('training time = {0:.2f} minuntes'.format(te/60))
```

Get Word Embeddings

In [ ]:
```
import pandas as pd
word_embed_layer = model.layers[2]
weights = word_embed_layer.get_weights()[0][1:]
print(weights.shape)
pd.DataFrame(weights, index=id2word.values()).head()
```

In [ ]:
```
from sklearn.metrics.pairwise import euclidean_distances
```

In [ ]:
```
distance_matrix = euclidean_distances(weights[0:5000])
print(distance_matrix.shape)
```

In [ ]:
```
similar_words = {search_term: [id2word[idx]
for idx in distance_matrix[word2id[search_term]-1].argsort()[1:6]+1]
    for search_term in [ ''لكره','لفرق' ,'لمغرب' ,'حكومه' ,'الفريق' ,'محمد','لطن', 'بن
لقدم' ,']}
similar_words
```

In [ ]:
```
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt
```

In [ ]:
```
words = sum([[k] + v for k, v in similar_words.items()], [])
words_ids = [word2id[w] for w in words]
word_vectors = np.array([weights[idx] for idx in words_ids])
print('Total words:', len(words), '\tWord Embedding shapes:', word_vectors.
shape)
tsne = TSNE(n_components=2, random_state=0, n_iter=10000, perplexity=3)
np.set_printoptions(suppress=True)
T = tsne.fit_transform(word_vectors)
labels = words
plt.figure(figsize=(14, 8))
plt.scatter(T[:, 0], T[:, 1], c='steelblue', edgecolors='k')
for label, x, y in zip(labels, T[:, 0], T[:, 1]):
    plt.annotate(label, xy=(x+1, y+1), xytext=(0, 0), textcoords='offset point
s')
plt.yscale('symlog')
plt.xscale('symlog')
```

In [ ]:
```
```

Unsupervised

In [ ]:
```
```

In [ ]:
```
# for printing and mapping back to the original text
totalvocab_stemmed = []
totalvocab_tokenized = []
for i in data['text'].head(10000):
      allwords_stemmed = (preprocessing_arabic(i))
      totalvocab_stemmed.append(allwords_stemmed)
      allwords_tokenized = nltk.word_tokenize(i)
      totalvocab_tokenized.append(allwords_tokenized)
```

In [ ]:
```
vocab_frame = pd.DataFrame({'words': totalvocab_tokenized}, index = totalvocab
_stemmed)
```

In [ ]:
```
vocab_frame.head(5)
```

In [ ]:
```
terms = tf_idf.get_feature_names()
```

In [ ]:
```
from sklearn.cluster import KMeans
```

In [ ]:
```
num_clusters = 5

km = KMeans(n_clusters=num_clusters)

%time km.fit(cv_unsupervised)

clusters = km.labels_.tolist()
```

In [ ]:
```

```

In [ ]:
```
frame = pd.DataFrame({'text': data['text'].head(10000),  'cluster': clusters})
```

In [ ]:
```
grouped = frame['text'].groupby(frame['cluster'])

grouped.count()
```

In [ ]:
```
from __future__ import print_function

print("Top terms per cluster:")
print()
order_centroids = km.cluster_centers_.argsort()[:, ::-1]
for i in range(num_clusters):
    print("Cluster %d words:" % i, end='')
    try:
         for ind in order_centroids[i, :5]:
            print(' %s' % terms[ind].split(' '))
    except Exception as e:
        pass
        print(e)
#     print()
#     print("Cluster %d words:" % i, end='')
#     for title in frame.ix[i]:
#         print(' %s,' % title, end='')
    print()
    print()
```

In [449]:
```
data.head(50)
```

Out[449]:

	text	label
0	...خوض المنتخب الوطني داخل القاعه مباراتين اعدادي	4
1	...مكنت مصالح الدرك الملكي لطنجه والعراءش ايقاف ر	1
2	...علنت نقابه الاتحاد الوطني للشغل بالمغرب الذراع	3
3	...خبارنا المغربيه سناء الوردي عرت صحيفه القدس ال	2
4	...مكنت المصالح الجمركيه عصر اليوم الجمعه بمعبر ب	1
5	...نسبه للجزاءريين المقيمين بالمغرب والمغاربه الم	3
6	...ينما الحارس الشاب يصارع لكي يستعيد عافيته كريم	4
7	...ادت السلطات المحليه لاقليم شفشاون بان سكان دوا	1
8	... لخياط يعتزم الترشح رغم افتقاده الشرط القانوني	4
9	...علم الجامعه الملكيه المغربيه لكره المدرب الاسب	4
10	...علنت المديريه العامه للامن الوطني العمليات الا	1
11	...خبارنا المغربيه يستعد رجل الاعمال احمد ابو هشي	0
13	...وفي شخصان واصيب اخرون بجروح متفاوته الخطوره مس	1
14	... دوه صحافيه الثلاثاء والمنتخب الغابوني يحل عبر	4
15	...فيلم الايراني ينزل القاعات السينماءيه ظل معارض	0
16	...عتصم طالب الاحزاب بالتوافق القضايا الجوهريه وا	3
17	... ثلت امام انظار وكيل الملك بالمحكمه الابتداءيه	1
18	... علم الصباح الرياضي سمير الزكرومي ومصطفي مراني	4
19	...خبارنا المغربيه الوزاني استضافت قناه فرانس امس	0
20	...شكل مسار الاقتصاد المغربي والرهانات الحاليه وا	2
21	...حمد الحليمي العلمي المندوب السامي للتخطيط ان ا	3
22	...شرع المكتب المسير لفريق الدفاع الجديدي تفعيل ا	4
23	... خصص المكتب السياسي لحزب التجمع الوطني للاحرار	3
24	...جل المكتب المسير لحسنيه اكادير لكره القدم جمعه	4
25	...زيد الطاقه الاستيعابيه للملعب بالفي متفرج وتخص	4
26	...طلقت نهايه الاسبوع الماضي بشاطء فم لبير بالداخ	4
28	...فظت تلميذه بالثانويه التاهيليه الفتح بمنطقه ال	1
29	... خبارنا المغربيه تابعه طالب زكرياء مومني البطل	3
30	...الق لافت للمغربي حمزه الساخي رفقه المنتخب الوط	4
31	...شرعت شركه الطيران الهولنديه كوريندون دوتش الخم	2
32	... حوار يوميه اخبار اليوم المغربيه نفت نجمه اراب	0
33	...فاقيه تمكن الشبكه بث اهم الانتاجات ومسءولون يت	0
34	... جهزا زوجه احدهما الحامل وطفليها والقيا بجثثهم	1
35	...سفرت قرعه كاس العالم بروسيا مواجهات ذات خصوصيه	4
36	...رور اسبوع الهجمات باريس مسرحا وراح ضحيتها قتيل	3

	text	label
37	...شار المهاجم الفرنسي كيليان مبابي انضم مءخرا لص	4
38	...انه سيكتفي بالاخراج وسيستعين بوجوه جديده العمل	0
39	...طوت الغرفه الجنحيه التلبسيه بالمحكمه الابتداءي	4
40	...خبارنا المغربيه قالت يوميه المساء عددها لنهايه	3
41	...علمت هسبريس مصدر مطلع عناصر القوات المسلحه الم	1
42	...خبارنا المغربيه سناء الوردي حوالي اسابيع قضاها	3
43	...خبارنا المغربيه سناء الوردي فضيحه بكل المقاييس	2
44	...جري وزير الشءون الخارجيه والتعاون السيد صلاح ا	3
45	...شف عبد اللطيف بروحو عضو لجنه الاقتصاد والماليه	2
46	...وسكير ان فيديو فتاه الطوبيس يعكس ازمه اخلاق ال	0
47	...مكن البطل المغربي سباق السيارات مهدي بناني بود	4
48	...خبارنا المغربيه متابعه قرر حزب التجمع الوطني ل	3
49	...راجعت تنازلها ومحاميها يكشف الاسباب ومبلغ التع	1
50	...عب نفي والنيابه العامه الاسبانيه تتهمه بالتهرب	4
51	...ز اتحاد طنجه الرجاء الجديدي بهدف لصفر امس الار	4

In [450]:
```
frame.head(50)
```

Out[450]:

	text	cluster
0	...خوض المنتخب الوطني داخل القاعه مباراتين اعدادي	4
1	...مكنت مصالح الدرك الملكي لطنجه والعراءش ايقاف ر	3
2	...علنت نقابه الاتحاد الوطني للشغل بالمغرب الذراع	0
3	...خبارنا المغربيه سناء الوردي عرت صحيفه القدس ال	3
4	...مكنت المصالح الجمركيه عصر اليوم الجمعه بمعبر ب	2
5	...نسبه للجزاءريين المقيمين بالمغرب والمغاربه الم	3
6	...ينما الحارس الشاب يصارع لكي يستعيد عافيته كريم	4
7	...ادت السلطات المحليه لاقليم شفشاون بان سكان دوا	2
8	... لخياط يعتزم الترشح رغم افتقاده الشرط القانوني	3
9	...علم الجامعه الملكيه المغربيه لكره المدرب الاسب	4
10	...علنت المديريه العامه للامن الوطني العمليات الا	2
11	...خبارنا المغربيه يستعد رجل الاعمال احمد ابو هشي	3
13	...وفي شخصان واصيب اخرون بجروح متفاوته الخطوره مس	3
14	... دوه صحافيه الثلاثاء والمنتخب الغابوني يحل عبر	4
15	...فيلم الايراني ينزل القاعات السينماءيه ظل معارض	3
16	...عتصم طالب الاحزاب بالتوافق القضايا الجوهريه وا	3
17	... ثلت امام انظار وكيل الملك بالمحكمه الابتداءيه	3
18	... علم الصباح الرياضي سمير الزكرومي ومصطفي مراني	1
19	...خبارنا المغربيه الوزاني استضافت قناه فرانس امس	3
20	...شكل مسار الاقتصاد المغربي والرهانات الحاليه وا	3
21	...حمد الحليمي العلمي المندوب السامي للتخطيط ان ا	3
22	...شرع المكتب المسير لفريق الدفاع الجديدي تفعيل ا	1
23	... خصص المكتب السياسي لحزب التجمع الوطني للاحرار	3
24	...جل المكتب المسير لحسنيه اكادير لكره القدم جمعه	3
25	...زيد الطاقه الاستيعابيه للملعب بالفي متفرج وتخص	1
26	...طلقت نهايه الاسبوع الماضي بشاطء فم لبير بالداخ	3
28	...فظت تلميذه بالثانويه التاهيليه الفتح بمنطقه ال	3
29	... خبارنا المغربيه تابعه طالب زكرياء مومني البطل	3
30	...الق لافت للمغربي حمزه الساخي رفقه المنتخب الوط	4
31	...شرعت شركه الطيران الهولنديه كوريندون دوتش الخم	3
32	... حوار يوميه اخبار اليوم المغربيه نفت نجمه اراب	3
33	...فاقيه تمكن الشبكه بث اهم الانتاجات ومسءولون يت	3
34	... جهزا زوجه احدهما الحامل وطفليها والقيا بجثثهم	2
35	...سفرت قرعه كاس العالم بروسيا مواجهات ذات خصوصيه	4
36	...رور اسبوع الهجمات باريس مسرحا وراح ضحيتها قتيل	3

	text	cluster
37	...شار المهاجم الفرنسي كيليان مبابي انضم مءخرا لص	3
38	...انه سيكتفي بالاخراج وسيستعين بوجوه جديده العمل	3
39	...طوت الغرفه الجنحيه التلبسيه بالمحكمه الابتداءي	1
40	...خبارنا المغربيه قالت يوميه المساء عددها لنهايه	0
41	...علمت هسبريس مصدر مطلع عناصر القوات المسلحه الم	3
42	...خبارنا المغربيه سناء الوردي حوالي اسابيع قضاها	3
43	...خبارنا المغربيه سناء الوردي فضيحه بكل المقاييس	3
44	...جري وزير الشءون الخارجيه والتعاون السيد صلاح ا	3
45	...شف عبد اللطيف بروحو عضو لجنه الاقتصاد والماليه	0
46	...وسكير ان فيديو فتاه الطوبيس يعكس ازمه اخلاق ال	3
47	...مكن البطل المغربي سباق السيارات مهدي بناني بود	3
48	...خبارنا المغربيه متابعه قرر حزب التجمع الوطني ل	0
49	...راجعت تنازلها ومحاميها يكشف الاسباب ومبلغ التع	3
50	...عب نفي والنيابه العامه الاسبانيه تتهمه بالتهرب	2
51	...ز اتحاد طنجه الرجاء الجديدي بهدف لصفر امس الار	1

In [ ]:

Supervised Learning

A- Naive Bayes Classifier

As a first step, we should find the most common words for each category. To do this, for each label, a list is created containing all the most frquent words throughout the documents. So, we will have 5 lists for 5 labels. Next, for each document, the cosine similarity sparse matrix will be generated and the index of the largest element in the matrix will be the corresponding label of the document. The same procedure is performed on the test data to find the labels of its texts.

1) Feature Engineering

In [477]:
```
vocab_frame['label']=list(data['label'].head(10000))
```

In [554]:
```
vocab_frame.head(2)
```

Out[554]:

	words	label
4	خوض,] المنتخب, الوطني, داخل, القاعه, ...,مباراتين	خوض المنتخب الوطني داخل القاعه مباراتين اعداديتين امام نظيره المصري يوليوز الجاري اكادير علي هامش المعسكر التدريبي يدخله اسود القاعه تحضيرا لنهاءيات كاس العالم بكولومبيا شتنبر المقبل وتاتي المباراتان الاعداديتان سياق الوقوف علي جاهزيه اللاعبين لمونديال كولومبيا خاصه المنتخب المصري يعد ابرز منافسي المنتخب الوطني القاره الافريقيه وسبق للمنتخب الوطني فاز علي نظيره المصري بثلاثه اهداف لاثنين نهاءي كاس افريقيا جنوب افريقيا وانهي المنتخب الوطني معسكره التدريبي طنجه الاربعاء الماضي بمشاركه العديد المحترفين ممن يمارسون الدوري الاوربي بعدما وجهت اليهم الجامعه الدعوه للوقوف علي جاهزيتهم هشام الدكيك مدرب المنتخب الوطني معسكر طنجه خاص بالمحترفين والهدف منه اختبارهم فنيا وبدنيا الحسم مدي اهليتهم وتابع اخترنا المحترفين اضافه اربعه محليين ممن يحظوا بفرصه اظهار مءهلاتهم الدكيك قرر اقامه معسكر تدريبي منطقه سوس يحظي المنتخب الوطني داخل القاعه بشعبيه كبيره مشيرا الي ستتخلله مباراتان اعداديتان امام مصر وكشف الدكيك سيختتم استعداداته لمونديال كولومبيا افران الظروف مواتيه لضمان تحضير جيد وزاد قاءلا سنستغل عامل الارتفاع سطح البحر الاعداد الجيد وبالتالي التاقلم اجواء ومناخ كولومبيا الشبيه بافران واوضح الدكيك المنتخب الوطني سيخوض مباراه اعداديه دوليه امام المنتخب البرتغالي الاحتكاك اكثر بالمدرسه الاوربيه عيسي الكامح
1	مكنت,] مصالح, الدرك, الملكي, لطنجه, والعراءش, ...	كنت مصالح الدرك الملكي لطنجه والعراءش ايقاف رقيب درك بمركز درك تطفت الموقوف ضبط متلبسا بتلقي مبلغ مالي قيمته عشره الاف درهم طرف تجار المخدرات والذي موضوع عده مذكرات بحث وطنيه وتم وضع المعني رهن تدابير الحراسه النظريه لتقديمه الي العداله فور انتهاء التحقيق مع

In [535]:
```
all_words0=[]
all_words1=[]
all_words2=[]
all_words3=[]
all_words4=[]
```

In [540]:
```
data_label_0=vocab_frame[vocab_frame['label']==0]
data_label_1=vocab_frame[vocab_frame['label']==1]
data_label_2=vocab_frame[vocab_frame['label']==2]
data_label_3=vocab_frame[vocab_frame['label']==3]
data_label_4=vocab_frame[vocab_frame['label']==4]
```

In [541]:
```
for lis1 in data_label_0['words'] :
    all_words0=all_words0+lis1
for lis1 in data_label_1['words'] :
    all_words1=all_words1+lis1
for lis1 in data_label_2['words'] :
    all_words2=all_words2+lis1
for lis1 in data_label_3['words'] :
    all_words3=all_words3+lis1
for lis1 in data_label_4['words'] :
    all_words4=all_words4+lis1
```

In [543]:
```
mostfreq_0=nltk.FreqDist(all_words0)
mostfreq_1=nltk.FreqDist(all_words1)
mostfreq_2=nltk.FreqDist(all_words2)
mostfreq_3=nltk.FreqDist(all_words3)
mostfreq_4=nltk.FreqDist(all_words4)
```

In [550]:
```
nb_of_top_words=1000
```

In [595]:
```
feature_label0=' '.join(list(mostfreq_0)[:nb_of_top_words])
feature_label1=' '.join(list(mostfreq_1)[:nb_of_top_words])
feature_label2=' '.join(list(mostfreq_2)[:nb_of_top_words])
feature_label3=' '. join(list(mostfreq_3)[:nb_of_top_words])
feature_label4=' '. join(list(mostfreq_4)[:nb_of_top_words])
```

2) Train the Model

In [559]:
```
count_vectorizer = CountVectorizer()
```

In [634]:
```
sparse_matrix = count_vectorizer.fit_transform([feature_label0,feature_label1,
feature_label2,feature_label3,feature_label4,data['text'][24]])
trial=list(cosine_similarity(sparse_matrix)[5,:5])
print(trial)
trial.index(max(trial))
```

```
[0.03454943755742926, 0.05177225431377269, 0.06043121679314222, 0.06902967241
836358, 0.2821537400523394]
```

Out[634]:
```
4
```

In [635]:
```
def model_bayes(text):
    sparse_matrix = count_vectorizer.fit_transform([feature_label0,feature_lab
el1,feature_label2,feature_label3,feature_label4,text])
    sim=list(cosine_similarity(sparse_matrix)[5,:5])
    return(sim.index(max(sim)))
```

In [636]:
```
training_pred=[]
for text in data['text'].head(10000):
    training_pred.append(model_bayes(text))
```

In Sample Error

In [639]:
```
in_sample_error = accuracy_score(training_pred, data['label'].head(10000))
```

In [660]:
```
print('The in accuracy score is',in_sample_error*100,"%")
```

```
The in accuracy score is 82.36 %
```

Prediction - Bayes

In [646]:
```
data_test=pd.read_csv(r'arabic_classification_test.csv')
```

In [650]:
```
data_test=data_test.dropna()
```

In [651]:
```
test_pred=[]
for text in data_test['text'].head(10000):
    test_pred.append(model_bayes(text))
```

Out of Sample Error

In [657]:
```
outof_sample_error = accuracy_score(test_pred, data_test['label'].head(10000))
```

In [658]:
```
print('The out of sample error is',outof_sample_error*100,"%")
```

```
The out of sample error is 78.88 %
```

B- SVM

In [661]:
```
cv = CountVectorizer(binary=False, min_df=0.0, max_df=1.0)
cv_train_features = cv.fit_transform(data['text'].head(10000))

# transform test articles into features
cv_test_features = cv.transform(data_test['text'].head(10000))
```

In [663]:
```
train_label_names=data['label'].head(10000)
test_label_names=data_test['label'].head(10000)
```

In [664]:
```
from sklearn.svm import LinearSVC
svm = LinearSVC(penalty='l2', C=1, random_state=42)
svm.fit(cv_train_features, train_label_names)
svm_bow_cv_scores = cross_val_score(svm, cv_train_features, train_label_names,
cv=5)
svm_bow_cv_mean_score = np.mean(svm_bow_cv_scores)
print('CV Accuracy (5-fold):', svm_bow_cv_scores)
print('Mean CV Accuracy:', svm_bow_cv_mean_score)
svm_bow_test_score = svm.score(cv_test_features, test_label_names)
print('Test Accuracy:', svm_bow_test_score)
```

```
C:\Users\user\Anaconda3\lib\site-packages\sklearn\svm\base.py:929: Convergenc
eWarning: Liblinear failed to converge, increase the number of iterations.
  "the number of iterations.", ConvergenceWarning)
C:\Users\user\Anaconda3\lib\site-packages\sklearn\svm\base.py:929: Convergenc
eWarning: Liblinear failed to converge, increase the number of iterations.
  "the number of iterations.", ConvergenceWarning)

CV Accuracy (5-fold): [0.94005994 0.92353823 0.9265     0.93246623 0.9254254
3]
Mean CV Accuracy: 0.9295979658972963
Test Accuracy: 0.8589
```

C- SVM with Stochastic Gradient Descent

In [665]:
```
from sklearn.linear_model import SGDClassifier
svm_sgd = SGDClassifier(loss='hinge', penalty='l2', max_iter=5, random_state=4
2)
svm_sgd.fit(cv_train_features, train_label_names)
svmsgd_bow_cv_scores = cross_val_score(svm_sgd, cv_train_features, train_label
_names, cv=5)
svmsgd_bow_cv_mean_score = np.mean(svmsgd_bow_cv_scores)
print('CV Accuracy (5-fold):', svmsgd_bow_cv_scores)
print('Mean CV Accuracy:', svmsgd_bow_cv_mean_score)
svmsgd_bow_test_score = svm_sgd.score(cv_test_features, test_label_names)
print('Test Accuracy:', svmsgd_bow_test_score)
```

```
C:\Users\user\Anaconda3\lib\site-packages\sklearn\linear_model\stochastic_gra
dient.py:561: ConvergenceWarning: Maximum number of iteration reached before
convergence. Consider increasing max_iter to improve the fit.
  ConvergenceWarning)
C:\Users\user\Anaconda3\lib\site-packages\sklearn\linear_model\stochastic_gra
dient.py:561: ConvergenceWarning: Maximum number of iteration reached before
convergence. Consider increasing max_iter to improve the fit.
  ConvergenceWarning)

CV Accuracy (5-fold): [0.93956044 0.93453273 0.932      0.93996998 0.9334334
3]
Mean CV Accuracy: 0.9358993183239106
Test Accuracy: 0.8897
```

D- Random Forest

In [666]:
```
from sklearn.ensemble import RandomForestClassifier
rfc = RandomForestClassifier(n_estimators=10, random_state=42)
rfc.fit(cv_train_features, train_label_names)
rfc_bow_cv_scores = cross_val_score(rfc, cv_train_features, train_label_names,
cv=5)
rfc_bow_cv_mean_score = np.mean(rfc_bow_cv_scores)
print('CV Accuracy (5-fold):', rfc_bow_cv_scores)
print('Mean CV Accuracy:', rfc_bow_cv_mean_score)
rfc_bow_test_score = rfc.score(cv_test_features, test_label_names)
print('Test Accuracy:', rfc_bow_test_score)
```

```
CV Accuracy (5-fold): [0.88611389 0.89455272 0.8815     0.88194097 0.8858858
9]
Mean CV Accuracy: 0.8859986932246391
Test Accuracy: 0.8549
```

E- Gradient Boosting

In [668]:
```
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier(n_estimators=10, random_state=42)
gbc.fit(cv_train_features, train_label_names)
gbc_bow_cv_scores = cross_val_score(gbc, cv_train_features, train_label_names,
cv=5)
gbc_bow_cv_mean_score = np.mean(gbc_bow_cv_scores)
print('CV Accuracy (5-fold):', gbc_bow_cv_scores)
print('Mean CV Accuracy:', gbc_bow_cv_mean_score)
gbc_bow_test_score = gbc.score(cv_test_features, test_label_names)
print('Test Accuracy:', gbc_bow_test_score)
```

```
CV Accuracy (5-fold): [0.7982018  0.79510245 0.795      0.8034017  0.8288288
3]
Mean CV Accuracy: 0.8041069553313328
Test Accuracy: 0.6665
```

In [ ]:
```

```