In [ ]:

```
import nltk
nltk.download('punkt')
import csv
```

Read a dataset that contains news_headlines for clustering

In [50]:

```
from pyspark.sql import Row
from pyspark.sql import SparkSession
from pyspark.ml.feature import HashingTF, IDF, Tokenizer
spark = SparkSession.builder.getOrCreate()
data = spark.read.format("csv").option("header",True).option("inferSchem
a",True).load("abcnews-date-text.csv")
data.show()
data = data.limit(500000)
```

```
+------------+--------------------+
|publish_date|       headline_text|
+------------+--------------------+
|    20030219|aba decides again...|
|    20030219|act fire witnesse...|
|    20030219|a g calls for inf...|
|    20030219|air nz staff in a...|
|    20030219|air nz strike to ...|
|    20030219|ambitious olsson ...|
|    20030219|antic delighted w...|
|    20030219|aussie qualifier ...|
|    20030219|aust addresses un...|
|    20030219|australia is lock...|
|    20030219|australia to cont...|
|    20030219|barca take record...|
|    20030219|bathhouse plans m...|
|    20030219|big hopes for lau...|
|    20030219|big plan to boost...|
|    20030219|blizzard buries u...|
|    20030219|brigadier dismiss...|
|    20030219|british combat tr...|
|    20030219|bryant leads lake...|
|    20030219|bushfire victims ...|
+------------+--------------------+
only showing top 20 rows
```

In [51]:

```
tokenizer = Tokenizer(inputCol="headline_text", outputCol="words")
wordsData = tokenizer.transform(data)
wordsData.show()
wordsData.count()
```

```
+------------+--------------------+--------------------+
|publish_date|       headline_text|               words|
+------------+--------------------+--------------------+
|    20030219|aba decides again...|[aba, decides, ag...|
|    20030219|act fire witnesse...|[act, fire, witne...|
|    20030219|a g calls for inf...|[a, g, calls, for...|
|    20030219|air nz staff in a...|[air, nz, staff, ...|
|    20030219|air nz strike to ...|[air, nz, strike,...|
|    20030219|ambitious olsson ...|[ambitious, olsso...|
|    20030219|antic delighted w...|[antic, delighted...|
|    20030219|aussie qualifier ...|[aussie, qualifie...|
|    20030219|aust addresses un...|[aust, addresses,...|
|    20030219|australia is lock...|[australia, is, l...|
|    20030219|australia to cont...|[australia, to, c...|
|    20030219|barca take record...|[barca, take, rec...|
|    20030219|bathhouse plans m...|[bathhouse, plans...|
|    20030219|big hopes for lau...|[big, hopes, for,...|
|    20030219|big plan to boost...|[big, plan, to, b...|
|    20030219|blizzard buries u...|[blizzard, buries...|
|    20030219|brigadier dismiss...|[brigadier, dismi...|
|    20030219|british combat tr...|[british, combat,...|
|    20030219|bryant leads lake...|[bryant, leads, l...|
|    20030219|bushfire victims ...|[bushfire, victim...|
+------------+--------------------+--------------------+
only showing top 20 rows
```

Out[51]:

```
500000
```

TF-IDF

In [52]:

```
hashingTF = HashingTF(inputCol="words", outputCol="rawFeatures",numFeatu
res=300)
featurizedData = hashingTF.transform(wordsData)
```

```
In [53]: featurizedData.show()
```

```
+------------+--------------------+--------------------+---------------
-----+
|publish_date|       headline_text|               words|         rawFea
tures|
+------------+--------------------+--------------------+---------------
-----+
|    20030219|aba decides again...|[aba, decides, ag...|(300,[42,57,12
2,1...|
|    20030219|act fire witnesse...|[act, fire, witne...|(300,[23,43,72,
11...|
|    20030219|a g calls for inf...|[a, g, calls, for...|(300,[46,66,14
4,1...|
|    20030219|air nz staff in a...|[air, nz, staff, ...|(300,[83,116,11
7,...|
|    20030219|air nz strike to ...|[air, nz, strike,...|(300,[45,88,10
9,1...|
|    20030219|ambitious olsson ...|[ambitious, olsso...|(300,[18,110,19
2,...|
|    20030219|antic delighted w...|[antic, delighted...|(300,[50,81,10
5,1...|
|    20030219|aussie qualifier ...|[aussie, qualifie...|(300,[0,39,51,1
10...|
|    20030219|aust addresses un...|[aust, addresses,...|(300,[79,83,11
1,1...|
|    20030219|australia is lock...|[australia, is, l...|(300,[2,62,100,
18...|
|    20030219|australia to cont...|[australia, to, c...|(300,[2,42,88,1
11...|
|    20030219|barca take record...|[barca, take, rec...|(300,[81,95,13
8,1...|
|    20030219|bathhouse plans m...|[bathhouse, plans...|(300,[107,139,1
97...|
|    20030219|big hopes for lau...|[big, hopes, for,...|(300,[44,168,22
5,...|
|    20030219|big plan to boost...|[big, plan, to, b...|(300,[88,123,20
4,...|
|    20030219|blizzard buries u...|[blizzard, buries...|(300,[11,105,15
4,...|
|    20030219|brigadier dismiss...|[brigadier, dismi...|(300,[35,129,19
5,...|
|    20030219|british combat tr...|[british, combat,...|(300,[35,95,11
7,1...|
|    20030219|bryant leads lake...|[bryant, leads, l...|(300,[84,88,20
0,2...|
|    20030219|bushfire victims ...|[bushfire, victim...|(300,[15,88,15
8,1...|
+------------+--------------------+--------------------+---------------
-----+
only showing top 20 rows
```

IDF features

In [54]:

```
idf = IDF(inputCol="rawFeatures", outputCol="features")
idfModel = idf.fit(featurizedData)
rescaledData = idfModel.transform(featurizedData)

rescaledData.select("headline_text", "features").show()
rescaledData = rescaledData.select("features")
```

```
+--------------------+--------------------+
|       headline_text|            features|
+--------------------+--------------------+
|aba decides again...|(300,[42,57,122,1...|
|act fire witnesse...|(300,[23,43,72,11...|
|a g calls for inf...|(300,[46,66,144,1...|
|air nz staff in a...|(300,[83,116,117,...|
|air nz strike to ...|(300,[45,88,109,1...|
|ambitious olsson ...|(300,[18,110,192,...|
|antic delighted w...|(300,[50,81,105,1...|
|aussie qualifier ...|(300,[0,39,51,110...|
|aust addresses un...|(300,[79,83,111,1...|
|australia is lock...|(300,[2,62,100,18...|
|australia to cont...|(300,[2,42,88,111...|
|barca take record...|(300,[81,95,138,1...|
|bathhouse plans m...|(300,[107,139,197...|
|big hopes for lau...|(300,[44,168,225,...|
|big plan to boost...|(300,[88,123,204,...|
|blizzard buries u...|(300,[11,105,154,...|
|brigadier dismiss...|(300,[35,129,195,...|
|british combat tr...|(300,[35,95,117,1...|
|bryant leads lake...|(300,[84,88,200,2...|
|bushfire victims ...|(300,[15,88,158,1...|
+--------------------+--------------------+
only showing top 20 rows
```

I only use the headline of news and that maybe the reason why the Error is so huge

In [59]:

```
from pyspark.ml.clustering import KMeans

kmeans = KMeans().setK(300).setSeed(1)
model = kmeans.fit(rescaledData)



wssse = model.computeCost(rescaledData)
print("Within Set Sum of Squared Errors = " + str(wssse))



centers = model.clusterCenters()
```

```
Within Set Sum of Squared Errors = 33112470.9293
```