Kyle Hundman
Text Analytics Assignment 5

1.

For this step, I included **194** countries and **202** capitals.

I chose to only include countries that were bolded and not italicized because the other 'countries' were dependent territories, overseas departments, or regions (not internationally recognized countries). Also, often these 'regions' had unofficial or disputed capitals when they had a capital at all.

For the internationally recognized countries, there were 7 countries with multiple capitals, all of which I indexed. I indexed all of them because they would help train the naïve Bayes classifier later in the assignment. The additional information will help make the classification model more accurate (distinguish better).

2.

Greek and Roman, not Persion:

Result count: 25
['Tripoli']
['Tunis']
['Sofia']
['Lisbon']
['Nicosia']
['Monaco']
['Skopje']
['Cairo']
['Ljubljana']
['Bucharest']
['Podgorica']
['Montevideo']
['Algiers']
['Bangui']
['Bern']
['Budapest']
['Bratislava']
['Berlin']
['Amsterdam']
['Madrid']
['Warsaw']
['Havana']
['Copenhagen']
['Buenos Aires']
['Abidjan']

Shakespeare:

Result count: 4
max_distance = 4, prefix = 2

['London']
['Prague']
['Cairo']
['Washington, D.C.']

'Located below sea level"

Result count: 1
['Baku']

3.

Top 30 Words for each continent (Antarctica has no internationally recognized countries)

AFRICA	ASIA	EUROPE	NORTH AMERICA	SOUTH AMERICA	OCEANIA
1 city	1 city	1 retrieved	1 city	1 city	1 retrieved
2 retrieved	2 retrieved	2 city	2 retrieved	2 retrieved	2 new
3 national	3 national	3 also	3 national	3 aires	3 islands
4 international	4 also	4 world	4 san	4 national	4 national
5 also	5 new	5 may	5 also	5 del	5 city
6 africa	6 international	6 national	6 central	6 buenos	6 canberra
7 south	7 world	7 london	7 mexico	7 san	7 zealand
8 african	8 university	8 international	8 district	8 main	8 wellington
9 central	9 capital	9 july	9 spanish	9 world	9 australian
10 cape	10 district	10 main	10 area	10 also	10 government
11 town	11 main	11 population	11 international	11 spanish	11 may
12 page	12 government	12 european	12 new	12 international	12 south
13 climate	13 may	13 first	13 united	13 central	13 capital
14 main	14 area	14 new	14 government	14 november	14 island
15 world	15 october	15 since	15 states	15 capital	15 original
16 area	16 population	16 june	16 world	16 area	16 act
17 capital	17 first	17 original	17 capital	17 montevideo	17 pacific
18 government	18 singapore	18 area	18 main	18 lima	18 september
19 may	19 asia	19 archived	19 citys	19 first	19 archived
20 french	20 east	20 largest	20 salvador	20 new	20 port
21 population	21 jerusalem	21 november	21 del	21 santiago	21 international
22 time	22 north	22 cities	22 washington	22 may	22 suva
23 cairo	23 cities	23 august	23 ottawa	23 government	23 april
24 new	24 since	24 see	24 american	24 america	24 october
25 country	25 south	25 museum	25 population	25 metropolitan	25 area
26 wikipedia	26 isbn	26 paris	26 public	26 district	26 honiara
27 include	27 september	27 january	27 park	27 cities	27 page
28 first	29 jakarta	28 centre	28 town	28 universidad	28 fiji
29 usage	29 history	29 athens	29 federal	29 park	29 also
30 october	30 march	30 copenhagen	30 first	30 bogotá	30 time

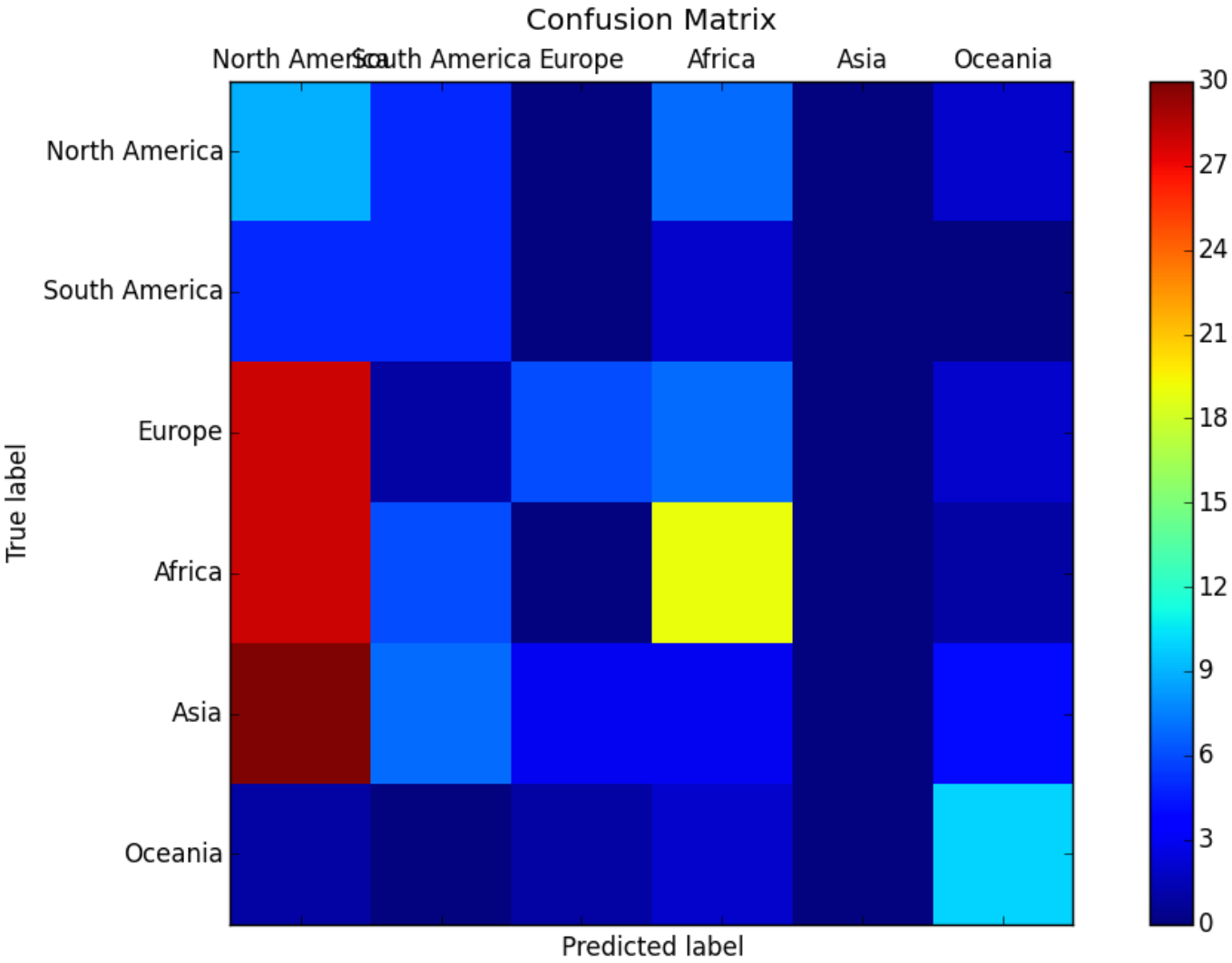
Unique Top30 Terms:

Count: **62** of 180

List:

	Unique Top Word	Continent	Unique Top Word	Continent
1	'athens'	'Europe'	'african'	'Africa'
2	'wellington'	'Oceania'	'port'	'Oceania'
3	'asia'	'Asia'	'european'	'Europe'
4	'april'	'Oceania'	'zealand'	'Oceania'
5	'ottawa'	'North America'	'australian'	'Oceania'
6	'include'	'Africa'	'university'	'Asia'
7	'largest'	'Europe'	'suva'	'Oceania'
8	'wikipedia'	'Africa'	'centre'	'Europe'
9	'island'	'Oceania'	'north'	'Asia'
10	'fiji'	'Oceania'	'states'	'North America'
11	'canberra'	'Oceania'	'east'	'Asia'
12	'climate'	'Africa'	'cairo'	'Africa'
13	'federal'	'North America'	'august'	'Europe'
14	'honiara'	'Oceania'	'country'	'Africa'
15	'isbn'	'Asia'	'history'	'Asia'
16	'salvador'	'North America'	'july'	'Europe'
17	'universidad'	'South America'	'act'	'Oceania'
18	'public'	'North America'	'singapore'	'Asia'
19	'pacific'	'Oceania'	'mexico'	'North America'
20	'aires'	'South America'	'time'	'Africa'
21	'citys'	'North America'	'america'	'South America'
22	'american'	'North America'	'paris'	'Europe'
23	'united'	'North America'	'cape'	'Africa'
24	'usage'	'Africa'	'french'	'Africa'
25	'africa'	'Africa'	'january'	'Europe'
26	'museum'	'Europe'	'washington'	'North America'
27	'lima'	'South America'	'montevideo'	'South America'
28	'buenos'	'South America'	'june'	'Europe'
29	'london'	'Europe'	'jerusalem'	'Asia'
30	'metropolitan'	'South America'	'santiago'	'South America'
31	'islands'	'Oceania'	'see'	'Europe'

Confusion Matrix:



Africa was the only class that was identified with mild success.

Favorite City:

Tangier

Classified as: **Europe**

Most Informative **Features**
contains(tunis) = False Europe : Africa = 1.1 : 1.0
contains(delhi) = False Europe : Asia = 1.1 : 1.0
contains(french) = False Europe : Africa = 1.1 : 1.0
contains(manila) = False Europe : Asia = 1.1 : 1.0
contains(gaborone) = False Europe : Africa = 1.1 : 1.0
contains(cairo) = False Europe : Africa = 1.1 : 1.0
contains(sudan) = False Europe : Africa = 1.1 : 1.0

I used textblog to build my naïve Bayes classifier, and by default textblog uses a very simplistic Boolean feature selection (whether a trained word is present or not). The above feature selection was looking for

the Asian and African words listed and couldn't find them, so it was eventually classified as Europe once this process was complete. In a real world implementation, custom features such as word structures or origin would be much more effective.

Probabilities for each continent:

Africa: 0.1
Europe: 0.22
Oceania: 0.22
Asia: 0.1
North America: 0.18
South America: 0.18