MSiA 490 Text Analytics HW3
Xiang Shawn Li

1 Boolean similarity
**********Most**********

1: bhargava--anuj lin--luis, similarity = 0.303230852558

2: sun--yuan tu--ye, similarity = 0.281988479136

3: bhargava--anuj bozoklar--berk, similarity = 0.281642037375

**********Least**********

1: leboeuf--mark mays--jacob, similarity = 0.0704594848625

2: herrera--alejandro leboeuf--mark, similarity = 0.0862581949178

3: maravilla--francisco nekkanti--siva-madhav, similarity = 0.0876267820585

2 tf-idf similarity
**********Most**********

1: bhargava--anuj lin--luis, similarity = 0.118973039632

2: cambo--scott heston--matthew, similarity = 0.103557395093

3: bhargava--anuj bozoklar--berk, similarity = 0.100004777132

**********Least**********

1: herrera--alejandro leboeuf--mark, similarity = 0.00591904709947

2: leboeuf--mark mays--jacob, similarity = 0.00717316748394

3: herrera--alejandro heston--matthew, similarity = 0.00866948761592

3 Indexing
Webscaping:
For special cases in which countries have more than one capital, I decided to only include one capital which contains one of “declared”, “official”, “legislative”, “constitutional” words (those key words are mutual exclusive). So it results 249 capital cities and 249 countries.

4 Retrieval

a. **'Greek' and 'Roman' but not 'Persian'**:

Query: +city_text:greek +city_text:roman -city_text:persian
Results found: 30
Tiraspol,Transnistria (0.16824351)
Sukhumi,Abkhazia (0.15047915)
Tunis,Tunisia (0.12993431)
Tripoli,Libya (0.11545949)
Lisbon,Portugal (0.11374697)
Sofia,Bulgaria (0.093560524)
Algiers,Algeria (0.090479545)
Bucharest,Romania (0.08921147)
Podgorica,Montenegro (0.08569528)
Cairo,Egypt (0.07498337)
Ljubljana,Slovenia (0.072851576)
Zagreb,Croatia (0.07191507)
Budapest,Hungary (0.066828944)
Bern,Switzerland (0.06499827)
Gibraltar,Gibraltar (0.06499827)
Monaco,Monaco (0.064668946)
Skopje,Republic of Macedonia (0.063978694)
Bangui,Central African Republic (0.060595714)
Montevideo,Uruguay (0.057751894)
Bratislava,Slovakia (0.056873485)
Berlin,Germany (0.055430524)
Madrid,Spain (0.050893016)
Dublin,Republic of Ireland (0.045446783)
Kiev,Ukraine (0.045446783)
Amsterdam,Netherlands (0.04304812)
Wellington,New Zealand (0.042496752)
Havana,Cuba (0.041530106)
Warsaw,Poland (0.04062392)
Copenhagen,Denmark (0.037872322)
Buenos Aires,Argentina (0.034608424)
performSearch done

b. **'Shakespeare'**:

Query: city_text:Shakespear~2
Results found: 4
London,United Kingdom (0.06778301)
Cairo,Egypt (0.06710176)
Prague,Czech Republic (0.06710176)
Washington, D.C.,United States (0.0575158)
performSearch done

c. **'located below sea level'**:

Query: city_text:"located below sea level"~10
Results found: 1
Baku,Azerbaijan (0.09373708)
performSearch done

d. interesting query: "Chinese And Food And Tea":

performSearch
Results found: 132
Ulaanbaatar,Mongolia (0.13469528)
Taipei,Taiwan (0.11757642)
Kathmandu,Nepal (0.10557398)
Nairobi,Kenya (0.08775205)
Damascus,Syria (0.060953166)
Manama,Bahrain (0.0592934)
Beijing,China (0.05603166)
Jakarta,Indonesia (0.055048224)
Singapore,Singapore (0.052300524)
Lilongwe,Malawi (0.048989665)
Hong Kong,Hong Kong (0.046609323)
Seoul,South Korea (0.04503538)
Port Louis,Mauritius (0.041688643)
Havana,Cuba (0.040883124)
Lima,Peru (0.03940561)
Manila,Philippines (0.03915231)
Yerevan,Armenia (0.03860923)
Saipan,Northern Mariana Islands (0.038289238)
Luanda,Angola (0.03624444)
Pyongyang,North Korea (0.034640927)
Suva,Fiji (0.034640927)
Bangkok,Thailand (0.031388607)
Hanoi,Vietnam (0.031377356)
Managua,Nicaragua (0.029698407)
Montevideo,Uruguay (0.02757802)
Dhaka,Bangladesh (0.025945174)
Bucharest,Romania (0.025577333)
Bamako,Mali (0.024494832)
Phnom Penh,Cambodia (0.023761097)
Port of Spain,Trinidad and Tobago (0.02266704)
Berlin,Germany (0.021978918)
San Juan, Puerto Rico,Puerto Rico (0.02114259)
Tashkent,Uzbekistan (0.021002043)
Caracas,Venezuela (0.021002043)
Accra,Ghana (0.020966865)
Adamstown, Pitcairn Island,Pitcairn Islands (0.02023762)

Tokyo,Japan (0.019595867)
Budapest,Hungary (0.019215588)
Kuala Lumpur,Malaysia (0.01885455)
Paramaribo,Suriname (0.01777624)
Rome,Italy (0.017320463)
Antananarivo,Madagascar (0.017146382)
Buenos Aires,Argentina (0.016785871)
Thimphu,Bhutan (0.015725149)
London,United Kingdom (0.014850686)
Monaco,Monaco (0.01470143)
Ankara,Turkey (0.01470143)
Paris,France (0.014652612)
Mexico City,Mexico (0.013289855)
Tegucigalpa,Honduras (0.012601226)
Castries,Saint Lucia (0.012569699)
Papeete,French Polynesia (0.012569699)
Dushanbe,Tajikistan (0.012569699)
Bandar Seri Begawan,Brunei (0.012443367)
Lusaka,Zambia (0.012443367)
Copenhagen,Denmark (0.012247416)
Gustavia, Saint Barthélemy,Saint Barthélemy (0.011805279)
Bishkek,Kyrgyzstan (0.010665744)
Madrid,Spain (0.0105010215)
Porto-Novo,Benin (0.01011881)
Sri Jayawardenepura Kotte,Sri Lanka (0.01005576)
Mogadishu,Somalia (0.00997689)
Stepanakert,Nagorno-Karabakh Republic (0.009540106)
Kingston, Jamaica,Jamaica (0.00888812)
Maseru,Lesotho (0.00879879)
Roseau,Dominica (0.00879879)
Juba,South Sudan (0.00879879)
Belgrade,Serbia (0.00879879)
Nouakchott,Mauritania (0.00879879)
Honiara,Solomon Islands (0.008708544)
Pago Pago,American Samoa (0.008432343)
Bujumbura,Burundi (0.008432343)
Conakry,Guinea (0.008432343)
Palikir,Federated States of Micronesia (0.008432343)
Khartoum,Sudan (0.008347593)
Brasília,Brazil (0.007542116)
Bangui,Central African Republic (0.007542116)
Georgetown, Guyana,Guyana (0.00754182)
San José, Costa Rica,Costa Rica (0.0071550794)
Maputo,Mozambique (0.007110496)
Basseterre,Saint Kitts and Nevis (0.006745874)
Ottawa,Canada (0.0062216837)
Tórshavn,Faroe Islands (0.0059026396)
Ouagadougou,Burkina Faso (0.0059026396)

San Salvador,El Salvador (0.005656587)
Amman,Jordan (0.005111836)
Rabat,Morocco (0.005059405)
Hargeisa,Somaliland (0.005059405)
Port Moresby,Papua New Guinea (0.005059405)
Windhoek,Namibia (0.005059405)
Monrovia,Liberia (0.005059405)
Dakar,Senegal (0.005059405)
Abuja,Nigeria (0.005059405)
Kinshasa,Democratic Republic of the Congo (0.005059405)
Santo Domingo,Dominican Republic (0.00502788)
Tirana,Albania (0.00502788)
Gibraltar,Gibraltar (0.00502788)
Minsk,Belarus (0.004770053)
Quito,Ecuador (0.004770053)
Abu Dhabi,United Arab Emirates (0.004770053)
performSearch done

5 Latent Dirichlet Allocation

10 most frequent topics and the 10 most frequent words in each of them:

Topic 8: from were which been after over has since also other
Topic 93: population one years about century three up 2009 than 10
Topic 85: country government economic largest countries per first new economy gdp
Topic 98: has its have national world international state which from education
Topic 11: between only high two mi known both large north many
Topic 77: from its his during which had he who two needed
Topic 55: government president were military forces war between many state republic
Topic 1: from first have government has other many which coast also
Topic 2: century most during period central music popular republic among power
Topic 12: country national river central south groups areas rural people country's

10 most frequent topics with counts

	Counts
8	225305
93	117417
85	102331
98	87161
11	85494
77	81170
55	72725
1	68452
2	45376

12 36952