Computational Journalism: Assignment 1

Pankhuri Kumar (pk2569)

AP ARTICLES

LSI Model

1. **Taxes and Market (and an uprising in Argentina):** ('-0.211*"tax" + -0.124*"income" + -0.121*"taxes" + 0.117*"bank" + -0.110*"average" + -0.108*"workers" + 0.103*"financial" + 0.099*"bush" + -0.098*"nikkei" + -0.098*"irs"')
2. **Religious tensions, Trade (and King George VI's coronation):** ('0.144*"kashmir" + 0.115*"pope" + 0.110*"tax" + 0.101*"share" + -0.101*"fire" + -0.100*"party" + 0.098*"india" + 0.097*"moslem" + 0.089*"jammu" + 0.089*"financial"')
3. **African aid and mixed:** ('0.105*"bill" + 0.091*"aid" + -0.090*"africa" + -0.090*"hospital" + -0.081*"ms" + -0.081*"smoking" + -0.078*"embassy" + -0.077*"keating" + 0.076*"korean" + -0.075*"sales"') – *not representative of the documents in this topic.*
4. **Unrest in China, union rights (and a faulty medical bill - 'mrs'):** ('-0.152*"mrs" + 0.151*"china" + 0.127*"chinese" + -0.121*"tax" + 0.116*"billion" + 0.114*"rights" + -0.106*"kashmir" + 0.106*"abortion" + -0.088*"ec" + 0.087*"dukakis"')
5. **Japanese-US, Germany-EU Trade relations, urban development, (and AIDS detection test):** ('-0.212*"german" + 0.190*"trade" + -0.166*"party" + -0.139*"east" + 0.128*"israel" + -0.124*"germany" + 0.121*"israeli" + -0.113*"west" + 0.107*"students" + 0.103*"japanese"')
6. **Financial Markets:** ('-0.234*"stock" + -0.211*"index" + -0.162*"dollar" + -0.148*"1" + -0.144*"0" + -0.144*"yen" + -0.144*"market" + -0.138*"prices" + -0.137*"rose" + -0.136*"percent"')
7. **Korea, students and earthquake:** ('-0.226*"korean" + -0.201*"students" + -0.190*"north" + -0.188*"korea" + 0.133*"mrs" + 0.130*"mandela" + -0.126*"school" + -0.125*"earthquake" + -0.115*"roh" + -0.112*"richter"')
8. **Flights, Japan, and IBM chips:** ('0.209*"computer" + -0.117*"air" + -0.101*"united" + 0.098*"bus" + -0.097*"police" + 0.096*"abortion" + -0.092*"states" + 0.087*"japanese" + -0.082*"chrysler" + 0.080*"bank"')
9. **Irish Republican Army & Britain, and a horrendous murder/suicide:** (51, '-0.133*"abortion" + 0.119*"irish" + 0.114*"ireland" + 0.111*"ira" + 0.109*"housing" + 0.104*"british" + 0.103*"jackson" + -0.102*"mrs" + 0.091*"northern" + -0.090*"women"')

10. **Trade and Germany:** ('-0.192*"dukakis" + -0.144*"company" + -0.143*"jackson" + 0.137*"german" + 0.123*"police" + -0.120*"south" + 0.119*"bush" + 0.117*"deficit" + -0.114*"corp" + 0.111*"trade"')

Above are the ten randomly sampled topics from the LSI model, for which I sampled 60 topics. I played around with the number of topics, but after about 50 topics, the newly generated topics seemed to be a repetition of another topic (or close enough to ignore). These topics do not seem to be entirely cohesive, rather they form a combination of various logical topics. For example, civil unrest in China is combined with union rights, which are two topics with similar words (rights, arrests, complaints), but in terms of news are quite far apart. Though LSI does a decent job of combining these up, they do not have much value in, let's say, an investigative project.

LDA Model

1. **Elections and voting, businesses by blacks, Denny's, and unions, :** ('0.005*"turnout" + 0.004*"arts" + 0.003*"municipal" + 0.003*"voter" + 0.002*"black" + 0.002*"133" + 0.002*"wrists" + 0.002*"grants" + 0.002*"novel" + 0.002*"milan"')
2. **Police charges, government fund-raising, hurricane:** ('0.004*"pacs" + 0.002*"dinner" + 0.002*"flooding" + 0.002*"island" + 0.002*"raising" + 0.002*"benson" + 0.002*"police," + 0.002*"parish" + 0.002*"sister" + 0.002*"casinos"')
3. **Flood, and mixed topics:** ('0.003*"royalties" + 0.002*"cafe" + 0.002*"compromise" + 0.002*"crandall" + 0.002*"balloon" + 0.001*"beachfront" + 0.001*"annie" + 0.001*"clouds" + 0.001*"$205" + 0.001*"``could"')
4. **A chemical accident, Unions and stock markets:** (46, '0.006*"stock" + 0.004*"listed" + 0.004*"shares" + 0.004*"index" + 0.003*"market" + 0.003*"common" + 0.003*"analysts" + 0.003*"points" + 0.002*"foam" + 0.002*"credit"')
5. **Drug dealers, dollar market values:** ('0.007*"dollar" + 0.006*"yen" + 0.006*"late" + 0.005*"bid" + 0.005*"canadian" + 0.004*"1" + 0.004*"compared" + 0.004*"dealers" + 0.004*"gold" + 0.004*"francs,"')

6. **Disruptive children, gun ownership:** ('0.004*"gun" + 0.003*"law" + 0.002*"boys" + 0.002*"creditors" + 0.002*"handguns" + 0.002*"maryland\'s" + 0.002*"children" + 0.002*"parents" + 0.002*"maryland" + 0.001*"cheap"')
7. **U.S. Factories, drug regulations, and Iraqi hostages:** ('0.005*"banning" + 0.003*"regulations" + 0.003*"safety" + 0.003*"address" + 0.002*"ives" + 0.002*"joints" + 0.002*"drug" + 0.002*"manufacturing" + 0.002*"industrial" + 0.002*"hoffman,"')
8. **Earthquakes, housing market:** ('0.004*"uss" + 0.004*"roberts" + 0.003*"angeles" + 0.003*"c" + 0.003*"los" + 0.003*"heat" + 0.003*"ago:" + 0.002*"lenders," + 0.002*"governors" + 0.002*"chinese"')
9. **Stock prices, rockets, and political unrest:** ('0.006*"electoral" + 0.003*"donald" + 0.003*"unchanged" + 0.003*"store" + 0.003*"up," + 0.003*"volume" + 0.002*"jones" + 0.002*"watch" + 0.002*"really" + 0.002*"book"')
10. **Heat wave, and flag desecration:** ('0.004*"filing" + 0.003*"tax" + 0.003*"history," + 0.003*"settled" + 0.003*"rostenkowski" + 0.003*"mock" + 0.002*"iwo" + 0.002*"discretionary" + 0.002*"boiling" + 0.002*"jima"')

LDA seems to do a better job of creating more cohesive topics. Though it still depends on common words, there seems to be overlap between the meaning of the topics as well. Further, LDA also seems to divide the articles more evenly among it's topics. In the LSI model, some topics would have only 2 or 3 articles but LDA seems to divide almost 20-30 articles for each of the topics I examined.

STATE OF THE UNION SPEECHES

It's easy to see a change in the topics of the State of the Union speeches over the decades of the 20th century. Certain words carry over for a few decades, before losing importance in the TF-IDF model. This is apparent with how the words "communism" and "soviets" appears with varying importance in 1950-1980. I tried to find the top-20 highest ranked words for each topic, and realized certain words (debt, commitment, tonight) turned up often only because of the intrinsic nature of a State of the Union speech. However, other words were more topical and could be connected to real-world events of the time.

The 1900s focus on the Panama Canal, and and forests and reserves, while the 1910s have TF-IDF terms that refer to the World War I, like submarines, cruisers, destroyers, prussian, scout.

The 1930s reflect the Depression, with words like depression, unemployment, banks and livelihood. The 1940s reflect the World War II, with axis, Hitler, nazis, japanese, 1945 being some of the main words in the corpus.

In a similar fashion, the 2000s spoke of Saddam Hussein, Iraq and terrorism. The 1990s had ‘21st’ as one of its highly ranked words, possibly alluding to the dawn of a new century. Soviet still made an appearance in the 1990s.

Overall, TF-IDF seems able to pull out the main themes of these speeches. With some better tokenization and lemmatization (I had Iraq and Iraqi show up together), I’m certain TF-IDF could attain even better results with this dataset.