**Clustering on Social Network**

Fjalet Kyqe: Data Mining ,Clustering , Social Network ,K-Means ,Word2Vec ,Elbow

Introduction

Rrjetet sociale kane perdorim shume te madh nga miliona përdorues në mbarë botën,një pjesë prej të cilëve janë edhe politikanë. Rrjetet sociale luaj role shume te madh ne shperndarje e lajme per nje kohe shume te shkurtur. Rrjetet sociale hapin horizonte të reja në cdo aspect të jetës dhe për këtë arsye përdoruesit mund të përfitojnë nga kjo mundësi. Gjithashtut ne rrjetet sociale kane perdorim edhe nga politikane te shume shteteve ne te cilin shprehin mendimet dhe lajmet rreth politikave shteterore , ekonomise ,luftrave etj.Prandaj në këtë punim qëllimi ynë është që të grumbullojmë tweet-at e politikanit më të njohur Presidentit te Shteteve të Bashkuara Donald Trump, si rezultat do të kemi disa grumbullime me disa fjalë kyqe të perafërta në bazë te shkrimeve të tweet-ave dhe ne do ta parashikojmë se cilat kanë qenë temat më të përfolura te Trump në llogarinë e tij të Twitter.

Dataset

Ky punim ka si per qellim clustering te fjaleve nga rrjetet sociale perkatesisht nga Twitter .Ne ne kete punim kemi perdorur nje dataset real ne te cilin jane marrur te gjithe tweet-at e Presidentit te Amerikes Donald Trump ne 5 vitet e fundit.Ne kete dataset pra kemi gjithsej 22819 tweets ne te cilat janw marrur keto te dhena :source -nga cila paisje eshte publikuar tweeti , text-teksit I shkruar nga Donald Trump , time-koha e publikimit , count\_favorite-numri i pelqimeve te tweetit ,count\_comments-numri i komenteve , is\_retweet-a eshte pergjigjur ne ndonje koment Trump . Dataseti eshte marrur nga kjo faqe [1].

2 Para-përpunimit ne data mining ne tekst

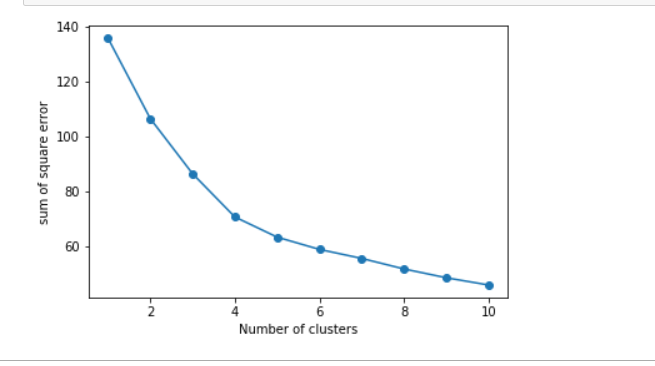
Gjatë procesit të mbledhjes së tekstit, teksti mund të jetë i organizuar lirshëm dhe mund të interpretohet si integrim i paarsyeshëm i tekstit ose mungesa e informacionit. Nëse teksti nuk është skanuar me kujdes për të identifikuar problemet , atëherë nxjerrja e tekstit mund të çojë në mos nxerrje te sakte te informative dhe informative te panevojshme. Teksti i pastrukturuar mund të çojë në analizë të dobët të tekstit që ndikon në saktësinë e një produkti . Faza e para-përpunimit organizon dokumentet në një numër fiks të kategorive të paracaktuara. Pra ne fazen e para-përpunimit hyne te gjitha ato hapa ne te cilen pastrohet teksti per clustering , disa prej tyre jane : hapsirat e panevojshme ,teksti ne shkronja te vogla ,heqja e shenjave te pikesimit ,heqja e stopwords , etj . Pasi para-përpunimit teksti eshte I gatshme per word2vec.

Word2Vec

Word2vec është një grup i modeleve të lidhura që përdoren për të prodhuar vektore nga një tekst me fjalë. Këto modele janë rrjete të cekëta me dy shtresa që janë trajnuar për rindërtimin e konteksteve gjuhësore të fjalëve. Word2vec merr si hyrje një përmbledhje të madhe të tekstit dhe prodhon një hapësirë vektoriale, zakonisht me disa qindra dimensione, me çdo fjalë unike në tekst duke u caktuar një vektor përkatës në hapësirë. Vektoret e fjalëve janë të pozicionuara në hapësirën vektoriale të tillë që fjalët që ndajnë kontekste të përbashkëta në tekst janë të vendosura në afërsi të njëri-tjetrit në hapësirë . Në këtë punim kemi perdorur Word2Vec ashtut që fjalët të naj kthej në vektorë 100 dimensional ,fjalët të cilat janë permendur me shumë se 50 here të i ruaj ne nje fjalore.

Elbow

Pas perdorimit te Word2Vec athere jeni të gatshme për të bërë Cluster pasi që tani fjalët i kemi shprehur në bazë te vektorëve , por si fillim per te caktuar numrin e Clusterve do ta perdorim metoden e Elbow me te cilin masin Sum of Square Error per disa clustera dhe paraqesim grafikun .Aty ku grafi nuk ka thyerje te madhe marrin numrin e clusterve.

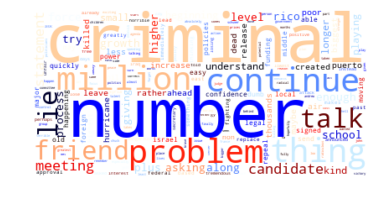


Shihet qarte se tek numri 4 i clusterve ka thyerje me shum ,prandaj do te bejme 4 cluster.

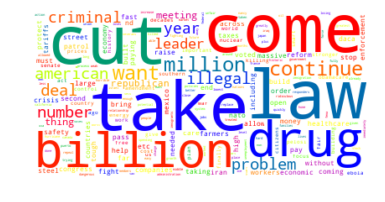
Clustering ne Word2Vec me Algoritmin K-Means

Pas punimit te literature survey qe kemi hulumtuar rreth Clustering ne Rrjete Sociale ,keni verejtur se Algoritmi K-means ka dalur me i perdorur ,por gjithashtut ka japur edhe rezultate te mira, andaj ne kete punim do te perdorim kete algoritem

Pas aplikimit te K-means do te shfaqen 4 cluster ne te cilin do te kene nga maximum 200 fjale per kluster ,nder te cilat do te jene ato fjale me te perdorurat



Cluster 1 :Nga keto fjale ne kete cluster mund te konkludojme se ka mbledhur tweets qe kane te bejne me tweets ne te cilin Trump ka folur per kriminelet ,problemet ,genjeshtrat e kanditateve tjere , humbje te milionave etj



Cluster 2: Nga keto fjale ne kete claster ,mund te konkludojme se ky cluster ka mbledhur tweets qe kane te bejne me:Ato tweets ne te cilen kane te bejne me gjerat ilegale qe behen ne amerike ,si te llojit te droges ,marrveshje te jashtligjme qe kane shkaktuar humbje me miliarda.etj

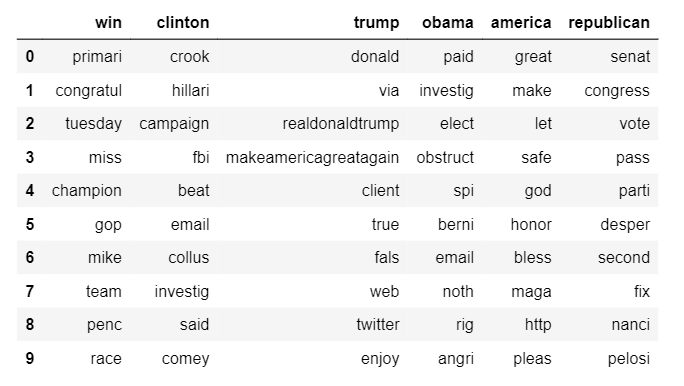


Cluster 3: Nga keto fjale ne kete claster ,mund te konkludojme se ky cluster ka mbledhur tweets qe kane te bejne me :deklaratat e kandidateve ,te thenat e tyre ,genjeshtat ,gjithashtut edhe fakte per ato.



Cluster 4: Nga keto fjale ne kete claster ,mund te konkludojme se ka mbledhur tweets qe kane te bejne me:deklaratat e tij ,punen qe ka bere ,tweets per shoket e tij ,udheheqje duke shprehur problemet me kriminel etj.

Pjese shtese e ketij projekti eshte kjo table ne te cilen me ane te algoritmin Word2Vec mund te gjejme ngjashmerite e disa fjaleve ,psh ne kemi marrur fjalet :win ,Clinton ,trump ,Obama ,America ,republican ,pastaj algoritmi ka gjetur 10 fjalet me te aferta me to.



References

[1] <http://www.trumptwitterarchive.com/archive>