Natural Language Processing

Assignment-2

Submitted by Ishaan Arora (201301011)

**Problem Statement**

We are given two datasets, English & Hindi. We are supposed to compute the co-occurrence matrix using top 250 words from corpus as feature words, then use K-means to cluster them into 50 groups & then display 25 words per group closest to the centroid.

For each dataset, we have to make co-occurence matrix for three cases:

1. Top 250 frequency ordered words including punctuations. All tokens exclding punctuations.

2. Top 250 frequency ordered words excluding punctuations and stop words. All tokens excluding punctuations.

3. Top 250 excluding top 50 frequency ordered words and excluding punctuations. All tokes excluding punctutations.

**Difficulties in K-means**

Since, it is a polynomial time algorithm, computational time was a major constraint. Since I had a deadline and my laptop does not have high computation power I had to constraint my iterations to 20. This made my co-occurence matrix sparse & most of the elements were zero.

In the intialisation of K centroids, I have chosen them randomly. If I would have somehow made some informed choice then my code wouldn't have taken that much time to run and hence the co-occurence matrix would have been small and compact

Sensitive to outliers.

OUTPUTS ATTAINED

These were the clusters obtained when the punctuations were not stripped.

Case 1 : stop words not included (some groups are mentioned here) (only some outputs displayed)

|  |
| --- |
| English , new , American , only , term , north , island , world , two , main , British , south , River , west , early , east , French , U.S. , band , largest , second , late , original , last , public |
| place , village , municipality |
| queen , distinction , restoration , valley , Gulf , Palace , leadership , truth , moment , sport , background , Commonwealth , flow , management , root , genus , parish , floor , spring , medium , union , monastery , picture , port , theme |
| sense , purpose , standard , value , cause , hands , President , course , concept , lack , location , range , command , centre , creation , source , president , Town , coast , production , majority , founder , center , nature , leader |
| such , known , well |
| be , this , have , point , go |
| queen , distinction , restoration , valley , Gulf , Palace , leadership , truth , moment , sport , background , Commonwealth , flow , management , root , genus , parish , floor , spring , medium , union , monastery , picture , port , theme |
| decision , history , role , major , people , career , war , film , way , long , great , works , system , period , government , common , popular , power , person , fam  ily , day , school , man , book , character |
| At , During , When , You , One , Its , Most , To , See , Some , Many , From , While , But , An , By , Although , These |
| 1927 , 1942 , 1924 , 1956 , 1941 , 1933 , 1930 , 1958 , 1938 , 1946 , 1929 , 1949 , 1934 , 1945 , 1918 , 1939 , 1917 , 1961 , 1935 , 1947 , 1955 , 1931 , 1937 , 1957 , 1922 |

Case 2 : stop words included

|  |
| --- |
| Harbor , Marathon , Richland , Burlington , Dodge , Fayette , Franklin , Dane , Polk , Collin , Cameron , Tarrant , Vernon , Shawano , Chippewa , Denton , Jefferson , Fairfax , Orange , Rusk , Wood , Harris , Marion , Monroe , Montgomery |
| During , Some , See , By , Since , With , Although , Other , To , One , Most , Many , From , While , But , An , These |
| also , has , are , that , at , from , they , Wyoming , though , USA , most , one , New , California , there , West , after , then , both , not , especially , along , so , Vermont |
| Kewaunee , Lexington , Hardin , Pennington , Wise , Okanogan , Ringkjøbing , Lafayette , Raleigh , Ribe , Ashland , Calumet , Klickitat , Vilas , Hutchinson , Weber , Logan , Benton , Shelby , Robertson , Knox , Mason , Claire , Carroll , Clay |
| Rosholt , Colton , Sturgis , Leola , Reliance , Okaton , Murdo , Wessington , Estelline , Headquartered , Roscoe , Tabor , Riverview , Jonesville , Pinewood , Millwood , Inman , Woodford , Branchville , Prosperity , Clarks , Elgin , Ridgeland , Edisto , Ridgeville |
| born , 2005 , now , pronounced , November , December , April , February , September , October , August , January , July , March , June , May , Liberal |
| 3,066. , 3,012. , 1,318. , 4,025. , 5,153. , 6,720. , 6,316. , 2,092. , 4,449. , 3,599. , 30,248. , 8,950. , 22,071. , 2,837. , 11,788. , 2,518. , 6,682. , 9,916. , 3,670. , 9,793. , 7,042. , 9,425. , 1,489. , 10,580. , 3,801. |
| languages , Empire , character , actor , theory , itself , Park , line , label , power , works , Rome , use , systems , success , Britain , water , law , Wales , Records , school , church , himself , period , state |
| \*Penstemon , Today , According , Born , Thus , Each , Later , \*The , More , # , Before , Two , Spoiler , Another , Neighboring , Like , Following , That , :For , Despite , All , Though , Because , Both , Among |
| before , no , since , more , if , even , United , named , Best , using , being , until , thus , whom , like , particularly , notably , during , due , according , making , based , among , having , under |

This is the output when punctuations were excluded (some outputs displayed)

Case 1 : stop words not included

|  |
| --- |
| growth , invention , reduction , Bank , conquest , lord , principles , strength , execution , expansion , fortress , passage , wing , Board , corner , shore , evolution , fragments , distribution , jurisdiction , doctrine , management , spirit , kings , department |
| ir , oldest , Latin , canal , greatest , summer , national , Russian , 19th , chief , stage , emperor , real , Scottish , royal , Canadian , European , surface , court , Japanese , Spanish , fourth , university , future , traditional |
| 1985 , 2001 , Lower , recent , 1987 , Japan , 1974 , 1978 , 1999 , 1996 , 1988 , 1977 , 1983 , 1984 , 1997 , 1986 , 1995 , 1998 , 1992 , 1980 , 1989 , 1994 , 1990 , 1991 , 1993 |
| ended , raised , playing , reached , discovered , run , carried , sold , provided , recorded , largely , mainly , primarily , captured , adopted , accepted , coined , marked , opened , employed , completed , kept , destroyed , invented , passed |
| Wise , Ribe , Harrison , Buffalo , Fairfield , Galveston , Marinette , Kenosha , Raleigh , Weber , Bennington , Whitman , Preston , Clay , Fremont , Wayne , Carroll , Lawrence , Augusta , Ashland , Crawford , Chesterfield , Lexington , Greenville , Lafayette |
| Rome , Canterbury , America , quality , Fame , Spain , William , George , Sir , Canada , Commons , Zealand , Castile , California , Sicily , Belgium , India , God , China , Austria , Sweden , Henry , Charles , Wales , Portugal |
| bishop , presence , method , care , branch , lack , nature , areas , piece , concept , origin , construction , species , command , History , source , possession , purpose , president , level , leader , director , loss , Department , value |
| user , genus , planet , TV , High , unitary , nation , screen , piano , media , ball , news , Nazi , cylinder , chemical , street , background , Mexican , poor , Portuguese , Danish , brain , federal , Welsh , cell |
| story , members , towns , version , king , family , parts , One , system , period , study , middle , case , beginning , history , daughter , side , author , Duke , capital , rest , development , House , Republic , head |
| to , as , and , at , with , on , for , by , from |

Case 2 : stop words included

|  |
| --- |
| only , other , most , album , band , French , work , Battle , word , film , movie , second , book , River , site , original , following , company , group , story , Beatles , battle , current , song , Twilight |
| Rutland , Elder , Paso , Spartanburg , Grayson , Oconto , Sheboygan , Portage , Addison , Skagit , Sumter , Harrison , Sweetwater , Washburn , Campbell , Navarro , Williamson , Trempealeau , Barron , Yakima , Burnett , Travis , Racine , Patricio , Orangeburg |
| Jutland , Lincoln , Grant , Taylor , Monroe , Polk , Rusk , Denton , Collin , Snohomish , Brown , Shawano , Pierce , Montgomery , Bergen , Hidalgo , Fairfax , Lake , Harris , Jackson , Wood , Marathon , Franklin , Jefferson , Cameron |
| beat , released , could , began , received , worked , retired , continued , appears , joined , attended , succeeded , never , found , did , graduated , subsequently , spent , remained , started , obtained , currently , gave , said , founded |
| Story , International , college , amount , Five , Next , remainder , oldest , Walt , Thin , Times , couple , Golden , resulting , lyrics , club , purpose , Red , exact , rest , building , Village , Right , Adventures , plot |
| at , addition , its , some , 2004 , October , November , 2003 , March , 2002 , fact , 2001 , order , May , 1996 , February , April , 1999 , July , January , December , August , September , June , mathematics |
| also , known , best , called , used , been , became , sometimes , well , better , commonly , often |
| Estates , Murphys , Ridgeville , Reevesville , Harleyville , Latta , Hartsville , NMPA , Motorsports , superspeedway , Lowrys , Richburg , Cheraw , McBee , Croghan , Pageland , Ruby , Manning , Paxville , Summerton , Turbeville , Cottageville , Edisto , Smoaks , Walterboro |
| winter , humans , authority , Street , conditions , Scotia , affairs , metals , policy , Westminster , troops , Austria , writing , poetry , programs , engines , problems , Constitution , Awards , Africa , education , Hungary , schools , Middle-earth , fiction |
| for , he , has , by , are , He , that , it , but , can , were , may , It , which , You |

OUTPUT FOR HINDI

Case 1 : when stop words aren't included

|  |
| --- |
| गूगल , तथापि , तीसरी , 19वीं , तभी , अर्थात् , चूँकि , कंपनी , सम्राट , नीम , मौर्य , प्रायः , मुगल , ऋग्वेद , 8 , जनवरी , मेरे , 9 , खुसरो , अर्थात , उदाहरणार्थ , मार्च , पृथ्वीराज , देहरादून , रामायण |
| लें , जीता , दे , चाहिए. , हाला , हुईं , होंगे , अपनाया , करो , करवाया , हुआ. , की. , जाये , निभाई , लगीं , करेंगे , है? , सकें , रहेगा , सकते , जाएगी , हूं , जायेगा , करे , करेगा |
| गांधी , आशा , शिव , महादेवी , कश्यप , गुप्त , साहिब , ब्रह्मा , शुक्ल , द्विवेदी , तुलसीदास , वशिष्ठ |
| भगवान , यदि , अपनी , उत्तर , राज्य , प्राचीन , इसकी , उसके , उसने , इनके , इनमें , इससे , वर्तमान , उन्हें , हालांकि , तब , आधुनिक , किंतु , इसे , उनका , अगर , यद्यपि , देश , सन , ऐसा |
| सन्‌ , वो , वास्तव , हम , 5 , यूरोपीय , इसलिये , २ , मंगल , समुद्र , इन्हें , द्वितीय , मनुष्य , आम , इन्हीं , वैसे , महाभारत , विकिपीडिया , मुंबई , 4 , अशोक , मानव , आजकल , पृथ्वी , अकबर |
| विजय , ज्ञान , स्थान , शिक्षा , अधिकार , जानकारी , डिग्री , मोक्ष , मान्यता , बहुमत , दर्जा , गौरव , सफलता , उपाधि , ख्याति |
| पाया , बोली , ले , मनाया , मानी , पाई , बनाई , समझा , आ , देखा , लिखी , गिना , लगाया , खेला , रह , मिल , बढ़ , जानी |
| लिखें , दीजिये , आयीं , डालें , दिलाया , निभाया , सम्भाला , बने. , करवाई , बिताया , नहीं. , रहो , डाले , जाओगे , बचाया , सौंपा , उभरा , फैलाया , बढाया , बढ़ी , लिखीं , मिला. , बढ़ा , रचा , रहना |
| उस , अपने , भारत , इसी , यहां , भारतीय , कुछ , उन्होंने , यहाँ , उनकी , सन् , वे , इसका , ये , जब , वह , उनके , इसमें , लेकिन |
| अधिक , करते , शामिल , कहते , होते , प्रसिद्ध , प्राप्त , कम , काफी , उपलब् |