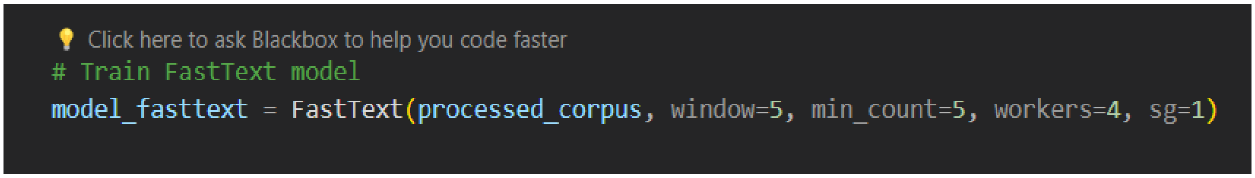
NLP Assignment 3

Ali Hisham Farouk 20200335

Sahar Hamdi 20201089

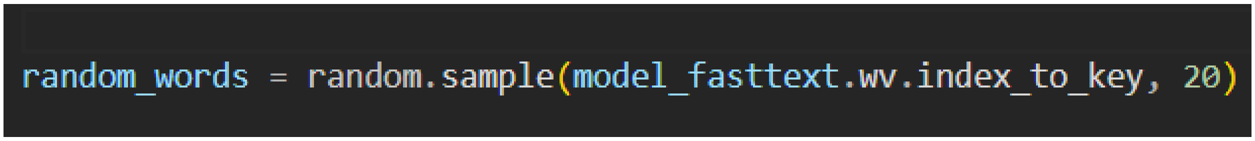
Hesham Mohamed Helmy 20200622

This is how we trained our model:



```
# Train FastText model
model_fasttext = FastText(processed_corpus, window=5, min_count=5, workers=4, sg=1)
```

Then we generated Random 20 words using the following line:



```
random_words = random.sample(model_fasttext.wv.index_to_key, 20)
```

The random words generated:

['amaretto', 'favs', 'lusherpride', 'torture', '810pm', 'filth', '5min', 'partition', 'save', 'panera', 'attacked', 'varying', 'limon', 'gras', 'reveal', 'multicolored', 'jockey', 'eaten', 'baja', 'mere']

First we tested them on our model, The results :

Word: amaretto

Closest: ['sorbetto', 'cioccolato', 'ganache', 'pistachio', 'marscapone', 'macadamia', 'meringue', 'creama', 'cardamom', 'crème']

Furthest: ['senior', 'inspected', 'inspect', 'safety', 'depart', 'inspector', 'departs', 'rate', 'inspection', 'resource']

Word: favs

Closest: ['faves', 'fav', 'favorite', 'fave', 'fava', 'favorito', 'favourite', 'favour', 'wordawesome', 'musthaves']

Furthest: ['urn', 'rm', 'enforce', 'enforced', 'urgency', 'emergency', 'violation', 'wage', 'internet', 'vaccine']

Word: lusherpride

Closest: ['guardiansofthegroove', 'phrasesfromplaces', 'nolalivemusic', 'dirtycoast', 'civicnola', 'algiersferry', 'uptownnola', 'lusher', 'thedandywarhols', 'nolaliving']

Furthest: ['accommodate', 'split', 'accommodated', 'carryout', 'ordering', 'amount', 'requested', 'request', 'offered', 'medium']

Word: torture

Closest: ['pasture', 'breathing', 'rapture', 'gesture', 'breather', 'snide', 'interruption', 'posture', 'cesspool', 'streamline']

Furthest: ['ri', 'fave', 'ni', 'ftw', '241', 'bogo', 'vodka', 'mf', 'oz', 'jai']

Word: 810pm

Closest: ['710pm', '910pm', '410pm', '510pm', '610pm', '210pm', '4pm10pm', '10pm', '8am10pm', '89pm']

Furthest: ['us', 'che', 'dal', 'texture', 'dent', 'lac', 'und', 'fake', 'undo', 'con']

Word: filth

Closest: ['filthy', 'urine', 'infested', 'disgusting', 'bedbug', 'disgust', 'disgusted', 'mop', 'disinfectant', 'unsanitary']

Furthest: ['monte', 'ita', 'frittata', 'cristo', 'empanadas', 'jo', 'pau', 'sammie', 'cravin', 'bentos']

Word: 5min

Closest: ['25min', '1015min', '15min', '45min', '10min', '30minute', '40min', '30min', '20minute', '30minutes']

Furthest: ['eco', 'petit', 'salvadoran', 'vodka', 'sante', 'inspired', 'yogurt', 'faves', 'flavor', 'ju']

Word: partition

Closest: ['demolition', 'tuition', 'ignition', 'proposition', 'petition', 'recognition', 'transition', 'audition', 'pollution', 'incarnation']

Furthest: ['tini', 'goood', 'gooood', 'goooood', 'gooooood', 'goooooood', 'gooooooood', 'goooooooood', 'goooo', 'yummm']

Word: save

Closest: ['saver', 'saved', 'saving', 'waste', 'money', 'wasted', 'savvy', 'wast', 'spend', 'adhere']

Furthest: ['hip', 'jai', 'ty', 'hello', 'ai', 'ken', 'ra', 'smokin', 'tai', 'ed']

Word: panera

Closest: ['paneras', 'pane', 'subway', 'starbuck', 'pdq', 'zaxbys', 'starbucks', 'publix', 'kroger', 'tjs']

Furthest: ['bud', 'led', 'sexy', '3d', 'scotch', 'hunt', 'tu', 'rare', 'tequila', 'tue']

Word: attacked

Closest: ['attack', 'attach', 'stacked', 'attachment', 'dialed', 'hacked', 'attached', 'whacked', 'jacked', 'racked']

Furthest: ['ba', 'tea', 'belgian', 'colombian', 'chai', 'sai', 'wine', 'baba', 'latte', 'tequila']

Word: varying

Closest: ['frying', 'implementing', 'heaping', 'qualifying', 'carrying', 'limiting', 'splendor', 'copying', 'various', 'varietal']

Furthest: ['ri', 'fuckin', 'fuck', 'lin', 'kann', 'sa', 'jos', 'ma', 'jenn', 'bobby']

Word: limon

Closest: ['limo', 'riquísimo', 'limonada', 'limpio', 'pésimo', 'benicio', 'cacio', 'mio', 'caldo', 'mauricio']

Furthest: ['urn', 'clothes', 'cloth', 'teenage', 'tend', 'ing', 'organized', 'teen', 'kid', 'nt']

Word: gras

Closest: ['mardi', 'mardigras', 'grasp', 'mardis', 'foie', 'mardigras2015', 'lundi', 'parade', 'zaras', 'madras']

Furthest: ['worked', 'worker', 'working', 'salsa', 'towel', 'asked', 'charge', 'asking', 'toned', 'questioned']

Word: reveal

Closest: ['reviver', 'rebekah', 'revel', 'solicitous', 'kunefe', 'affair', 'rev', 'revue', 'deidre', 'revival']

Furthest: ['dunkin', 'drivethru', 'starbucks', 'gas', 'bagel', 'wawa', 'cv', '247', 'mc', 'sbucks']

Word: multicolored

Closest: ['colored', 'discolored', 'multitask', 'multi', 'watercolor', 'cologne', 'allnatural', 'fermented', 'perplexed', 'coloring']

Furthest: ['parking', 'shuttle', 'close', 'parkin', 'thu', '1am', '2am', '5am', 'thurs', 'thur']

Word: jockey

Closest: ['hockey', 'sockeye', 'mickey', 'dickey', 'pacman', 'buckeye', 'hawkeye', 'hockessin', 'softball', 'lsu']

Furthest: ['medi', 'pupusas', 'brazilian', 'ethiopia', 'eta', 'cafe', 'flavour', 'rely', 'medium', 'latte']

Word: eaten

Closest: ['ive', 'weve', 'ate', 'tastiest', 'uneaten', 'craved', 'hadand', 'havent', 'gotten', 'saddest']

Furthest: ['pin', 'vin', 'ua', 'tu', 'lug', 'shirt', 'dora', 'sunglass', 'ski', 'jean']

Word: baja

Closest: ['raja', 'roja', 'tostados', 'adobada', 'sopes', 'asado', 'taquitos', 'guadalajara', 'quesabirria', 'pastor']

Furthest: ['lobby', 'older', 'ge', 'wireless', 'chest', '70', 'passage', 'pump', 'registration', 'massage']

Word: mere

Closest: ['chevre', 'bere', 'sucre', 'merit', 'chevelle', 'chouteau', 'cheshire', 'biere', 'deidre', 'meridien']

Furthest: ['ing', 'karaoke', 'in', 'fryer', 'wash', 'tsa', '11pm', '1pm', 'annoying', 'dryer']

Then we tested it on the pre trained model, The results:

Word: amaretto

Closest: ['Amaretto', 'Frangelico', 'frangelico', 'liqueur', 'kahlua', 'anisette', 'cointreau', 'amaretti', 'limoncello', 'Kahlua']

Furthest: ['SDMS', 'ASTRO', 'HealthWatch', 'ITEX', 'KAMS', 'SEWA', 'SEMI', 'IDSA', 'SGIA', 'SSTI']

Word: favs

Closest: ['faves', 'fav', 'fave', 'fav.', 'favs.', 'favorties', 'FAVE', 'favortie', 'fave.', 'favorites']

Furthest: ['..', '..', '..', '..', '..', '..', '...', '..', '...', '..']

Word: lusherpride

Closest: ['airplanealertapp-downloadarrow-downarrow-expandarrow-leftarrow-right-alternatearrow-rightarrow-upattractionbadgeballoonsbarcodebellbookbuildingcalendarcameracarcartcash-backchairchat-dotschatcheck-circlecheckboxclosecomputercreatecruiseemptybadgeenvelope-open-outlineenvelope-openenvelopefacebookfingerflowerfoodfootballgeargift-cardgiftgridguitarhealthheart-outlinehearthome-gardeninfojewellifetime-cashbacklipsticklocationlockmapmedalmenuminus-circlemoneynewsomnichannelpacifierpaperclippawphoneplayplugplusprintpromotedquestion-markrebatermn-rsearchshareshirtshoesliderstar-outlinestarstopwatchstoretag-addtagthumbs-downthumbs-uptoytrophyuserwatchx', 'DEky4M0BSpUOTPnSpkuL5I0GTSnRI4jMepcaFAoxIoFnX5kmJQk1aYvr2odGBAAIfkECQoABAAsCQAAABAAEgAACGcAARAYSLCgQQEABBokkFAhAQEQHQ4EMKCiQogRCVKsOOAiRocbLQ7EmJEhR4cfEWoUOTFhRIUNE44kGZOjSIQfG9rsyDCnzp0AaMYMyfNjS6JFZWpEKlDiUqALJ0KNatKmU4NDBwYEACH5BAkKAAQALAkAAAAQABIAAAhpAAEQGEiQIICDBAUgLEgAwICHAgkImBhxoMOHAyJOpGgQY8aBGxV2hJgwZMWLFTcCUIjwoEuLBym69PgxJMuDNAUqVDkz50qZLi', 'DEky4M0BSpUOTPnSpkuL5I0GTSnRI4jMepcaFAoxIoFnX5kmJQk1aYvr2odGBAAIfkECQoABAAsCQAAABAAEgAACGcAARAYSLCgQQEABBokkFAhAQEQHQ4EMKCiQogRCVKsOOAiRocbLQ7EmJEhR4c fEWoUOTFhRIUNE44kGZOjSIQfG9rsyDCnzp0AaMYMyfNjS6JFZWpEKlDiUqALJ0KNatKmU4NDBwYEACH5BAUKAAQALAkAAAAQABIAAAhpAAEQGEiQIICDBAUgLEgAwICHAgkImBhxoMOHAyJOpGgQY8aBGxV2hJgwZMWLFTcCUIjwoEuLBym69PgxJMuDNAUqVDkz50qZLi', '60-post-invoice-ninja-free-open-source-invoicing-amp-time-tracking-opposenewapstandardsus', 'GaelicSerbianSesothoShonaSindhiSinhalaSlovakSlovenianSomaliSpanishSudaneseSwahiliSwedishTajikTamilTeluguThaiTurkishUkrainianUrduUzbekVietnameseWelshXhosaYiddishYorubaZulu', 'crescendosexibloguerobateyabsorbersexiindesignabledinerolatifundiosexibrezarcularsutesexirapoplinbrezarcorrentosoVd.lazadareflejoreglafeministabrezarchuzasexiouttiquebloguerolin', '6-post-how-to-make-an-invoice-with-sample-invoices-wikihow-opposenewapstandardsus', 'ZJiUJNWmL69qHRgQACH5BAkKAAQALAkAAAAQABIAAAhnAAEQGEiwoEEBAAQaJJBQIQEBEB0OBDCgokKIEQlSrDjgIkaHGy0OxJiRIUeHHxFqFDkxYUSFDROOJBmTo0iEHxva7Mgwp86dAGjGDMnzY0uiRWVqRCpQ4lKgCydCjWrSplODQwcGBAAh', '60-post-invoice-ninja-free-open-source-invoicing-amp-time-tracking-lacey-chabertus', '60-post-invoice-ninja-free-open-source-invoicing-amp-time-tracking-ediblewildsus']

Furthest: ['ãŽã', '\U00100077', '\U00100071', '\U00100040', 'Cve', 'FollowPatrick', 'Shannonfrom', 'MattCAG', 'trse', '\U00100068']

Word: torture

Closest: ['torture.The', 'water-boarding', 'Torture', 'tortures', 'torture-', 'torturing', 'toture', 'waterboarding', 'torture.I', 'torturers']

Furthest: ['BusinessWest', 'Bryte', 'Muckey', 'TMCF', 'Cayzer', 'Beiträge', 'NetSpot', 'CCEDC', 'PlaneSense', 'EveryoneOn']

Word: 810pm

Closest: ['ShippingMJM', 'debloguerore flejoantecedentesexitlacuachebateysuteindesignableabsorbersexilatifundiosexibrezarsutemultiétnicosexiplinrapobrezarcorrentosoVd.lazadafisiochillidomabrezarsico-chuzaoutcolodrablogueroin', 'BLOGTWITTERYOUTUBEFACEBOOKTUMBLRTOONZONEFAQ', 'DEky4M0BSpUOTPnSpkuL5I0GTSnRI4jMepcaFAoxIoFnX5kmJQk1aYvr2odGBAAIfkECQoABAAsCQAAABAAEgAACGcAARAYSLCgQQEABBokkFAhAQEQHQ4EMKCiQogRCVKsOOAiRocbLQ7EmJEhR4cfEWoUOTFhRIUNE44kGZOjSIQfG9rsyDCnzp0AaMYMyfNjS6JFZWpEKlDiUqALJ0KNatKmU4NDBwYEACH5BAkKAAQALAkAAAAQABIAAAhpAAEQGEiQIICDBAUgLEgAwICHAgkImBhxoMOHAyJOpGgQY8aBGxV2hJgwZMWLFTcCUIjwoEuLBym69PgxJMuDNAUqVDkz50qZLi', 'DEky4M0BSpUOTPnSpkuL5I0GTSnRI4jMepcaFAoxIoFnX5kmJQk1aYvr2odGBAAIfkECQoABAAsCQAAABAAEgAACGcAARAYSLCgQQEABBokkFAhAQEQHQ4EMKCiQogRCVKsOOAiRocbLQ7EmJEhR4cfEWoUOTFhRIUNE44kGZOjSIQfG9rsyDCnzp0AaMYMyfNjS6JFZWpEKlDiUqALJ0KNatKmU4NDBwYEACH5BAUKAAQALAkAAAAQABIAAAhpAAEQGEiQIICDBAUgLEgAwICHAgkImBhxoMOHAyJOpGgQY8aBGxV2hJgwZMWLFTcCUIjwoEuLBym69PgxJMuDNAUqVDkz50qZLi', 'GaelicSerbianSesothoShonaSindhiSinhalaSlovakSlovenianSomaliSpanishSudaneseSwahiliSwedishTajikTamilTeluguThaiTurkishUkrainianUrduUzbekVietnameseWelshXhosaYiddishYorubaZulu', 'ZJiUJNWmL69qHRgQACH5BAkKAAQALAkAAAAQABIAAAhnAAEQGEiwoEEBAAQaJJBQIQEBEB0OBDCgokKIEQlSrDjgIkaHGy0OxJiRIUeHHxFqFDkxYUSFDROOJBmTo0iEHxva7Mgwp86dAGjGDMnzY0uiRWVqRCpQ4lKgCydCjWrSplODQwcGBAAh', 'HobbsSalterSamsungSataliteSeboSennheiserServisSharpSiemensSKYSmegSmilightSONSonorousSonosSonyStovesSwanSylvaniaTechlinkTefalTeknixTJ', '0e46e5e7c2a7aca07365ecb6ca1e5a9e', '81-post-18-free-service-invoice-templates-in-word-and-excel-hloomcom-cool-math-gamesus']

Furthest: ['\U00100071', '\U00100077', '201es', 'alsod', '\U00100040', '\U00100068', 'SpotlightCircaMy', 'accompagnés', '100x100px', 'Relationally']

Word: filth

Closest: ['filthy', 'foulness', 'excrement', 'dirtiness', 'vileness', 'squalor', 'cesspit', 'disgustingness', 'filthiness', 'ordure']

Furthest: ['Single-Sign-On', 'Auto-MDIX', 'TACC', 'ITCA', 'NSMS', 'MentorNet', 'PortalGuard', 'IdenTrust', 'z114', 'Eurocopter']

Word: 5min

Closest: ['10min', '15min', '2min', '3min', '5mins', '20min', '10mins', '7min', '5-10min', '10-15min']

Furthest: ['4.0For', 'Consorting', 'non-Wikipedia', 'Hussein-era', 'funeral-related', 'Surnamed', 'Fuson', 'OfficeSpeech', 'Noggle', 'yet-to-be-discovered']

Word: partition

Closest: ['partitions', 'partion', 'partitioning', 'Partition', 'parition', 'partiton', 'partitioned', 'partition.', 'partioning', 'partions']

Furthest: ['Erick', 'Ichthus', 'Tianxiao', 'Startt', 'Haggstrom', 'Nordqvist', '---Anonymous', 'Noochie', 'exaggerated.Here', 'Stian']

Word: save

Closest: ['saving', 'saved', 'saves', 'Save', 'save.', 'tosave', 'save.I', 'conserve', 'sacrificing', 'economize']

Furthest: ['IGEA', 'Recreative', 'NARB', 'XLR-11', 'AGCC', 'Markeri', 'DQP', 'LLCP', 'ISED', '1973a']

Word: panera

Closest: ['starbucks', 'quiznos', 'starbucks.', 'Panera', 'mcdonalds.', 'mcds', 'bread.', 'chickfila', 'wegmans', 'pizza.']

Furthest: ['Emitters', 'Reinsurers', 'Papapetrou', 'Phoria', 'post-2006', 'Awdry', 'Corfield', 'AutoView', 'Valuer', 'Rank-Broadley']

Word: attacked

Closest: ['assaulted', 'ambushed', 'atacked', 'attacking', 'attcked', 'chased', 'Attacked', 'menaced', 'attacked.', 'terrorized']

Furthest: ['moreProject', 'moreFull', 'StockOut', 'version1', '0Board', 'SCMG', 'version2', 'YesAdditional', 'moreRichard', 'moreRegister']

Word: varying

Closest: ['differing', 'varied', 'varing', 'Varying', 'varrying', 'various', 'different', 'ranging', 'widely-varying', 'differeing']

Furthest: ['VARCHARCan', 'hereMake', 'LAKings', 'LetsGo', 'dady', 'CodeGo', "'Let", 'MUMU', 'now.claiming.message', "'Take"]

Word: limon

Closest: ['limón', 'limonada', 'limone', 'jugo', 'clementina', 'limonum', 'limonade', 'pepino', 'aurantifolia', 'agua']

Furthest: ['PDSA', 'HomeFront', 'Apprentices', 'Commenting', 'HealthWatch', 'phase-out', 'ConservativeHome', '2011.If', 'control-oriented', 'CustomMade']

Word: gras

Closest: ['Gras', 'fois', 'veau', 'foie', "'etat", 'poivre', 'fraiche', 'résistance', 'etat', 'bouche']

Furthest: ['Upholder', 'OneCode', 'TaqMan', 'Guidepost', 'LifeSpan', 'JMW', 'article.--', 'NeatReceipts', 'CHRONOS', 'PayNearMe']

Word: reveal

Closest: ['revealing', 'reveals', 'revealed', 'uncover', 'reaveal', 'divulge', 'unveil', 'uncovers', 'conceal', 'divulged']

Furthest: ['Bussing', 'HardRock', 'FOCO', '57miles', 'KUPS', 'T.10', 'Liimatta', '12791', 'Symrna', '2.101']

Word: multicolored

Closest: ['multi-colored', 'multicoloured', 'multi-coloured', 'multi-hued', 'rainbow-colored', 'single-colored', 'brightly-colored', 'rainbow-hued', 'multi-patterned', 'multi-color']

Furthest: ['Kinbrace', 'fhall', 'Strensham', 'NSAI', 'Romsey', 'CIPFA', 'Imker', 'Terowie', 'Aftersales', 'Tonbridge']

Word: jockey

Closest: ['jockeys', 'Jockey', 'jocky', 'racehorse', 'ex-jockey', 'reinsman', 'owner-trainer', 'jockies', 'steeplechasers', 'money-winning']

Furthest: ['Jyllinge', 'Fantoft', 'Hjem', 'Norheimsund', 'UTSee', 'Skriv', 'CSNTM', 'MSEK', 'Velkommen', 'Aryl']

Word: eaten

Closest: ['ate', 'eatten', 'eaten.I', 'eaten.', 'eaten.The', 'eat', 'devoured', 'Eaten', 'consumed', 'munched']

Furthest: ['Clarifying', 'Relationship-based', 'Vergence', 'Three-step', 'FS6', 'Glidepath', 'DivisionNew', 'Geosolutions', 'InformationFeatures', 'Vizion']

Word: baja

Closest: ['bajas', 'Baja', 'ensenada', 'jalisco', 'sahara', 'BAJA', 'tecate', 'mexicali', 'caribe', 'cali']

Furthest: ['WeekCatawba', 'Bebris', 'Chamberlain', 'Peers', 'Stevereads', 'closed-doors', 'Harrow', 'theagency', 'Lords', 'Kalfin']

Word: mere

Closest: ['paltry', 'merely', 'measly', 'merest', 'mear', 'scant', 'miniscule', 'measley', 'trifling', 'puny']

Furthest: ['Rispin', 'IHSP', 'officiating.Burial', 'doWhich', 'Keyvelop', 'Greets', 'ResponsesAuthor', 'Selects', 'BRUBECK', 'answerWhich']

<u>Conclusion:</u>

The trained model operates better than the pretrained one, for example the pretrained model couldn't find the furthest of word favs and also for another example word 810pm bugged on both closest and furthest.