УДК 811.512.145

М.Р. Сәйхунов,
А.М. Хөсәенова,
Р.Р. Хөсәенов

МӘГЪНӘДӘШ СҮЗЛӘР ТАБУДА НЕЙРОН ЧЕЛТӘР МӨМКИНЛЕКЛӘРЕ

В работе рассматривается применение векторов слов, обученных на текстовом корпусе с помощью технологии на основе нейронных сетей word2vec, для поиска близких по значению слов, а также для решения аналогий вида «X – Y, Z – ?».

Ключевые слова: татарский язык, корпус, нейронные сети, word2vec, векторы слов.

The work discusses the use of word vectors trained on the text corpus using the word2vec technology based on neural networks to search for words that are close in meaning, as well as to solve analogies of the form «X – Y, Z – ?».

Keywords: Tatar language, corpus, neural networks, word2vec, word vectors.

Тел гыйлеме кысаларында нейрон челтәрләр. Соңгы елларда компьютер лингвистикасы өлкәсендә күпсанлы яңа технологияләр кулланыла башлады. Шуларның иң киң таралганнарыннан нейрон челтәрләрне атарга мөмкин. Аларның башка технологияләрдән төп аермасы, кагыйдәләр белән баетылуда, программалаштырылуда түгел, ә «өйрәтелү» принцибына корылуларында чагылыш таба. Шуңа күрә нейрон челтәрләр башта зур мәгълүмат базасында «өйрәтелә» һәм бу халәттә яңа мәгълүматны эшкәртә, аңа карата «фараз кыла» ала.

Әлеге эшкә нейрон челтәрләр ярдәмендә өйрәтелгән сүз векторлары җәлеп ителә, ягъни текстлардагы һәр сүзгә чын саннардан торган n-үлчәмле вектор (n-dimensional vector of real numbers) билгеләнә. Мәсәлән, әгәр n=3 булса, «*йорт*» сүзенә [0.5, 2.6, 1.5] кыяфәтендәге вектор туры килергә мөмкин. Сүзләрнең болай бирелеше лингвистиканың «Дистрибутив семантика» [Дистрибутивная семантика] өлкәсендә кулланыла. Моннан тыш сүз векторлары үзләрен телне табигый рәвештә эшкәртү [Natural language processing] өлкәсенә караган күпсанлы проблемаларны хәл итүдә дә нәтиҗәле күрсәттеләр (мәсәлән, машина тәрҗемәсе, текстның тематикасын яки тональлеген [Анализ тональности текста] билгеләү һ. б.).

Сүз векторларының үзенчәлеге һәм файдасы. Сүз векторлары идеясенең асылы «*охшаш тирәлектә очраган сүзләр якын мәгънәгә ия*» дигән гипотезага нигезләнә. Димәк текстларның зур күләмле тупланмасы булса, һәр сүзгә аерым вектор туры китерергә мөмкин. Якын мәгънәле сүзләргә туры килә торган әлеге векторлар n-үлчәмле пространствода бер-берсенә якын торырга тиеш булалар.

Алда әйтеп киткәнчә, соңгы вакытта бу мәсьәләне чишүдә нейрон челтәрләр кулланыла. Шуларның берсе, әлеге эштә карала торган word2vec моделе, сүзләрнең ихтималлы контекстын фаразларга өйрәтелгән гади нейрон челтәр аша сүз векторларын булдыра ала.

Шушы юл белән тупланган векторлар кайбер кызыклы һәм файдалы үзенчәлекләргә ия. Алар сүзләр арасындагы семантик һәм синтаксик мөнәсәбәтләрне таба һәм әлеге мөнәсәбәтләрнең ныклыгын билгели. Мондый мөмкинлекләрнең кайберләре алга таба 4 бүлектә карала.

Косинус охшашлыгы төшенчәсе. Оригиналь мәкаләдә word2vec моделендә ике төрле архитектура тәкъдим ителә [Tomas Mikolov]. Аларның берсендә (Skip-Gram) нейрон челтәр сүзнең ихтималлы контекстын фаразларга, икенчесендә (CBOW) контекст буенча ихтималлы сүзне ачыкларга өйрәтелә. Контекст бу очракта әлеге сүзнең уң һәм сул ягында килә торган һәм гадәттә уннан артмаган берничә күрше сүзне белдерә. Әлеге эштә без CBOW архитектурасын кулланабыз.

Word2vec векторлары рәвешендә бирелгән сүзләр арасындагы мөнәсәбәтләрне анализлау өчен, *косинус охшашлыгы* [Cosine similarity] төшенчәсенә мөрәҗәгать итәләр. Ул – векторлар арасындагы почмакның косинусын үлчәү аша ике векторның якынлыгын билгели торган күрсәткеч. Почмак кечерәк булган саен, векторлар бер-берсенә якынрак. Шулай итеп, әлеге почмакның косинусы 1 гә якынрак булып чыга.

Әгәр векторлар бер-берсенә каршы юнәлгән булса, косинус охшашлыгы -1 гә тигез булачак. Әгәр векторлар ортогональ булса, ягъни алар арасындагы почмак 90° тәшкил итсә, косинус охшашлыгы 0 гә тигез булачак. Әгәр векторлар бер тарафка юнәлгән булса, косинус охшашлыгы 1 гә тигез булачак. Шулай итеп, саннар арасындагы күрсәткечләр сүзләрнең төрле дәрәҗәдәге якынлыгы турында сөйли.

Әлеге күрсәткечне кулланып, аерым бер сүз векторына иң якын векторларны, ягъни күрше сүзләрне табарга (4.1 пункты), ике сүзнең охшашлык дәрәҗәсен бәяләргә (4.3 пункты) мөмкин. Бу күрсәткеч шулай ук 4.2 пунктындагы аналогияләрне төзүдә дә кулланыла, ләкин монда исәпләүләр бераз катлаулырак. Алар векторлар өстендәге арифметик гамәлләргә нигезләнә. Мәсәлән, *«җәй – эссе, ә кыш – ?»* аналогиясен алсак, *«?»* урынында без «*суык*» яки «*салкын*» кебек сүз булырга тиешлеген аңлыйбыз. Бу типтагы мәсьәләләрне автоматик рәвештә сүз векторлары аша чишеп була. Моның өчен без түбәндәге векторны исәпләп чыгара алабыз:

X = вектор(«*эссе*») – вектор(«*җәй*») + вектор(«*кыш*»).

Шуннан соң сүз векторлары пространствосында без X векторына иң якынын эзлибез дә, шуны җавап («*?*») буларак кайтарабыз. Әгәр сүз векторлары җитәрлек дәрәҗәдә яхшы өйрәтелгән булса, без дөрес җавапка ирешәчәкбез (бу очракта «*суык*» сүзе).

Word2vec моделен корпус нигезендә куллану. Word2vec моделе нигезендә сүз векторларын өйрәтү өчен аерым ике корпус алынды:

– күләме 15 миллион сүз тәшкил иткән һәм тулысынча матур әдәбият әсәрләренә корылган «Татар матур әдәбияты корпусы» [Татар матур әдәбияты корпусы];

– 356 миллион сүздән торган һәм үз эченә төрле жанрлардагы, төрле стильләрдәге текстларны алган «Татар теленең язма корпусы» [Татар теленең язма корпусы].

Корпуслардагы мәгълүмат ике формада әзерләнде: оригиналь (ягъни сүзформалар буларак) һәм леммалар рәвешендә. Башта автоматик рәвештә тотрыклы тезмәләр табылды. Шуннан соң әлеге сүзләр арасындагы мөнәсәбәтләр билгеләнде.

Исәпләүләрне башкару өчен, Google компаниясе тарафыннан эшләнгән word2vec программасы кулланылды [Word2vec].

Төзелгән система әлеге ике корпус сайтында да «Тезаурус» исеме белән урнаштырылды. Ул түбәндәге өч гамәлне башкарырга мөмкинлек бирә:

1. Мәгънәдәш сүзләр. Әлеге функция мәгънәдәш сүзләрне эзләргә мөмкинлек бирә. Мәсәлән, «*китап*» сүзенең косинус охшашлыгы буенча иң якын очраклары буларак түбәндәгеләр табыла. Мисаллардагы саннар һәр сүзнең «*китап*» сүзенә карата исәпләнгән косинус охшашлыгын күрсәтә:

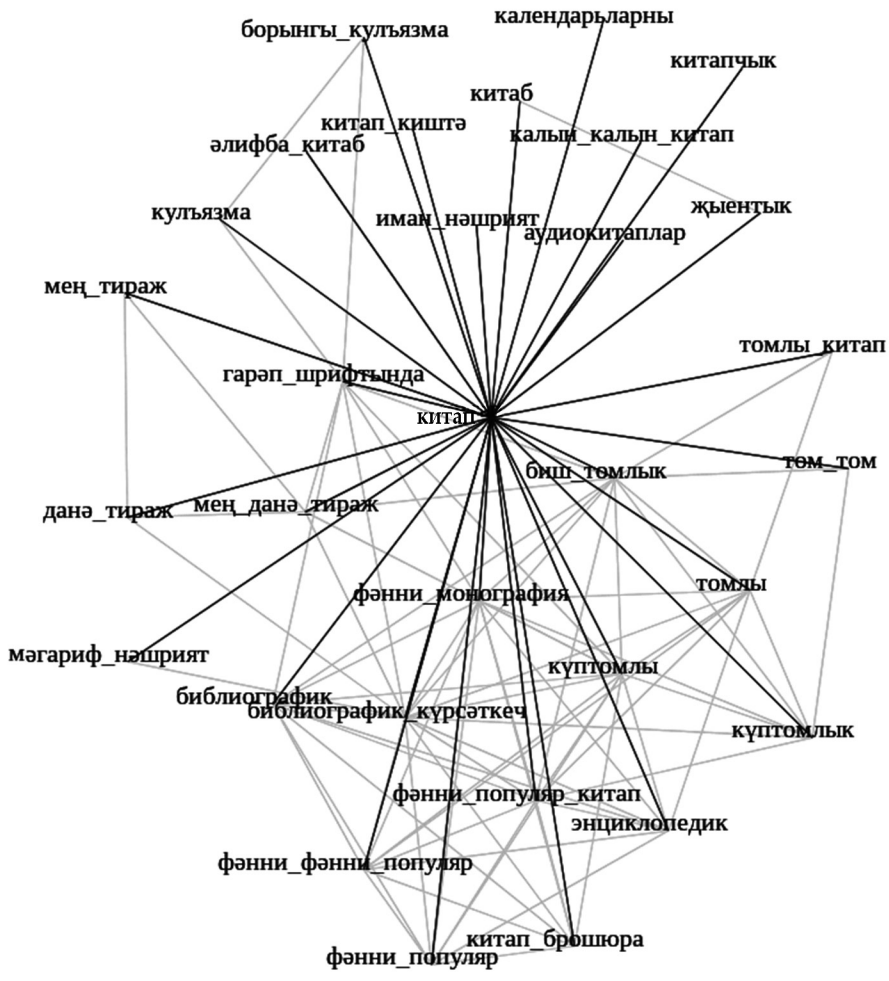
китап *– 0.752229 китаб, 0.682486 мең_данә_тираж, 0.672231 китапчык, 0.671023 том_том, 0.667688 фәнни_популяр_китап, 0.65399 китап_брошюра, 0.645721 иман_нәшрият, 0.643474 фәнни_популяр, 0.64313 җыентык, 0.634892 данә_тираж, 0.634619 энциклопедик, 0.634452 гарәп_шрифтында, 0.6336 әлифба_китаб, 0.6328 калын_калын_китап, 0.63182 библиографик_күрсәткеч, 0.631582 күптомлы, 0.63082 томлы, 0.628358 библиографик, 0.625116 календарьларны, 0.625025 томлы_китап, 0.624684 китап_киштә, 0.622028 борынгы_кулъязма, 0.620939 мәгариф_нәшрият, 0.620155 фәнни_монография, 0.619905 биш_томлык, 0.616826 кулъязма, 0.616399 аудиокитаплар, 0.616167 күптомлык, 0.613371 мең_тираж, 0.612622 фәнни_фәнни_популяр.*

Алдагы графикта табылган сүзләрнең «*китап*» сүзенә карата гына түгел, ә сыеграк төс белән бирелгән сызыклар аша үзара мөнәсәбәтләре дә чагылыш таба (1 нче график).

Икенче мисал буларак «кыр» сүзен карап үтик (2 нче график):

кыр *– 0.937906 көтүлек, 0.926447 чулман, 0.915009 көтү_көтү, 0.902445 итил, 0.901285 җәйләү, 0.901143 ерганак, 0.901018 җәйлә, 0.899437 буа, 0.899136 кышлау, 0.897992 аръягында, 0.895145 кырында, 0.894153 ялан, 0.893301 сөр, 0.887688 елга, 0.886665 үзән, 0.88585 чулман_буй, 0.883478 елга_буй, 0.88304 иген_кыр, 0.88128 киң_дала, 0.879818 алан, 0.879712 үзәнлек, 0.87778 каберлек, 0.877585 баткак, 0.876186 тикле, 0.875428 ташу, 0.875172 туздырып, 0.873972 җире, 0.873611 казы, 0.872944 нократ, 0.871216 тау_итәк.*

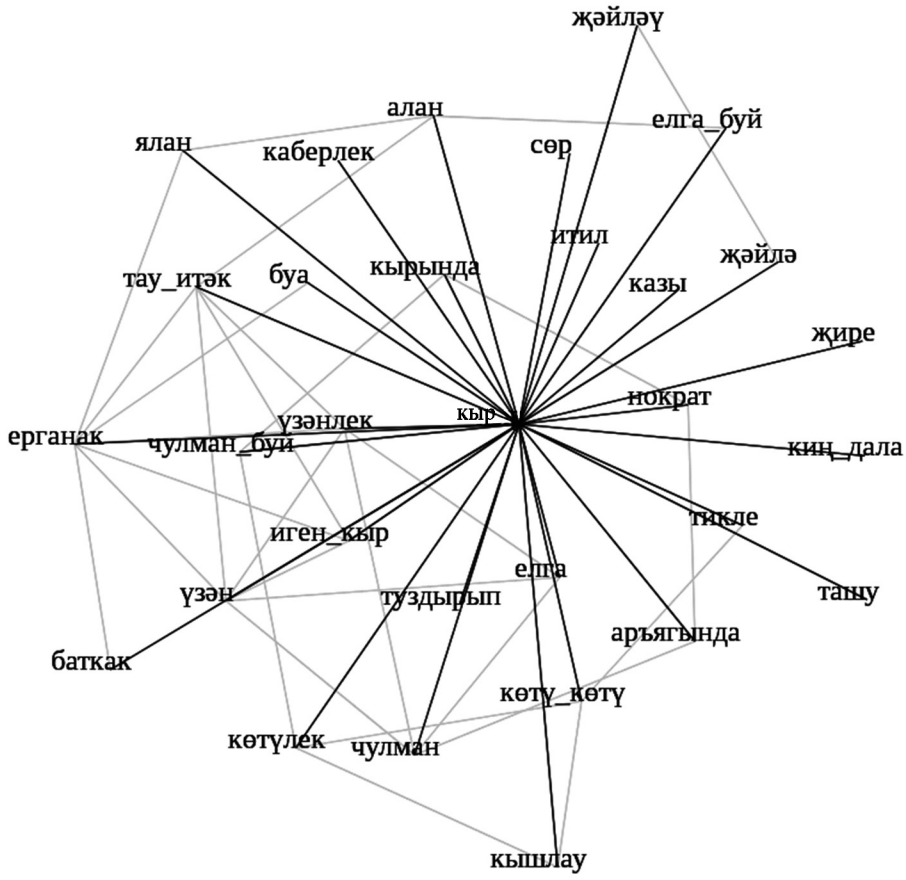
Түбәндәге мисалда берьюлы ике сүзгә карата уртак мәгънәдәш сүзләр табыла. Аның өчен ике сүзнең векторлары кушыла, һәм бу куш-



1 нче график

ма векторга иң якын сүз векторлары туплана (3 нче график). Шунысы кызык, табылган нәтиҗәләр исемлегенең башында әлеге ике сүз өчен дә уртак булган сыйфатларны (*«тимер»нең каты, ә «кәгазь»нең юка булуы*) чагылдырган сүзләр урнаша: *катыргы, картон, калай*:

тимер, кәгазь *– 0.611537 металл, 0.610027 катыргы, 0.609618 кәгазь_кисәк, 0.593739 картон, 0.590721 калай, 0.568972 тимер_кисәк, 0.567635 кәгазъ, 0.560675 арткы_яг, 0.560555 пластмасса, 0.557492 пластмасс, 0.556839 авторучка, 0.554691 тартмачык, 0.553912 ватман, 0.552176 карандаш, 0.551499 дүрт_бөклә, 0.550077 күкрәк_кесә, 0.546048 өстәл_тартма, 0.545314 сырлап, 0.54462 пыяла, 0.543587 акбур, 0.54297 тимерчыбык, 0.541497 мөһерлә, 0.540843 төргәклә, 0.537996 тартма, 0.536848 тоткыч, 0.535645 арматура, 0.534824 лупа, 0.534338 тәрәзә_рам, 0.533764 сырлы, 0.53317 конвертлар.*



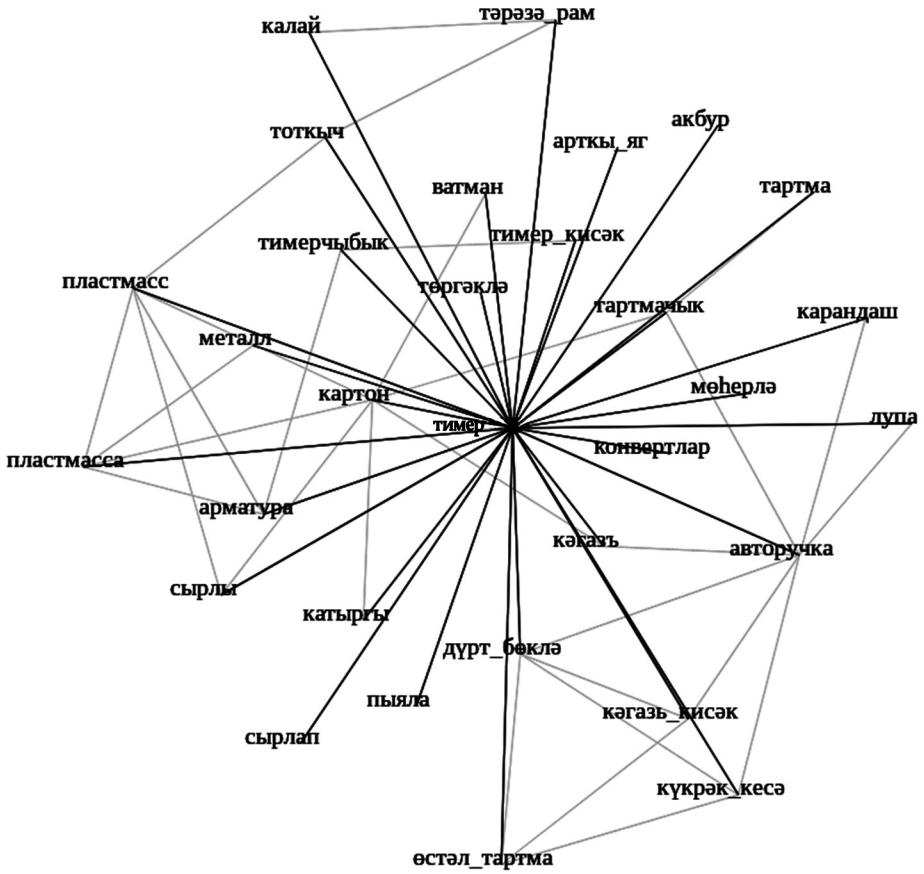
2 нче график

Ә «юрган, карават» пары өчен иң якын сүзләр буларак түбәндәгеләр табыла (4 нче график):

юрган, карават – *0.731421 сәке, 0.729877 йомшак_мендәр, 0.7238 одеял, 0.715946 кушетка, 0.715095 ятак, 0.711984 тимер_карават, 0.710837 ак_җәймә, 0.702655 ястык, 0.699629 яткырды, 0.698992 җәймә, 0.692825 мендәр, 0.688617 чыбылдык, 0.688601 раскладушка, 0.683847 карават_кырый, 0.682108 диван_карават, 0.680681 матрас, 0.673922 чишенде, 0.67388 түшәк, 0.672862 матрац, 0.667898 йомшак_түшәк, 0.666311 тәрәзә_пәрдә, 0.665535 түр_сәке, 0.665025 идән_келәм, 0.663881 тумбочка, 0.663878 шифоньер, 0.663257 кроватька, 0.660866 юрган_ябын, 0.660195 мамык_мендәр, 0.659965 тапчан, 0.659901 идән_палас.*

2. Төшенчәләр аналогиясе. Әлеге функция «*X – Y, Z – ?*» арифметик аналогиясе нигезендә нәтиҗәне табарга мөмкинлек бирә. Мәсәлән, «*эт – ул хайван, ә алма – ул нәрсә?*» һ. б. очракларны карап үтик:

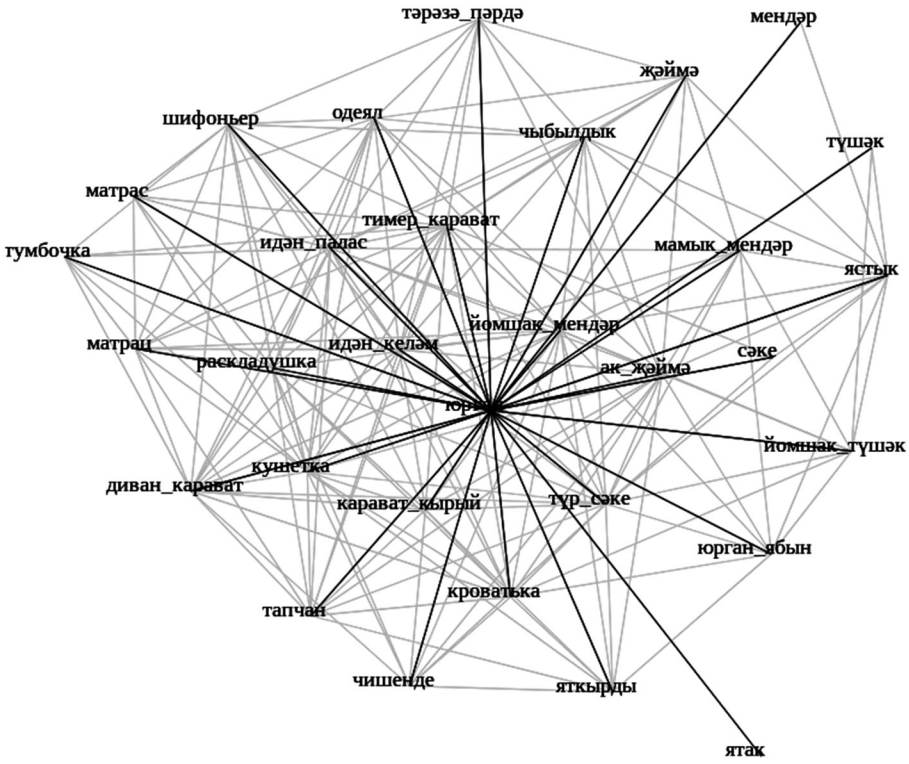
эт – хайван, алма – *0.873757 җимеш, 0.862394 гөл, 0.851239 үсемлек, 0.848565 шифалы, 0.847187 искитмәле, 0.834528 бәрәкәт-*



3 нче график

ле, 0.833419 күрке, 0.827192 чәчәк, 0.815759 татып_кара, 0.815128 пакь, 0.810416 чәчкә, 0.808018 гөлбакча, 0.80672 мул, 0.806341 тансык, 0.800148 йок, 0.798206 шытып_чык, 0.795633 орлык, 0.793072 сабак, 0.786472 хозурлык, 0.779775 тиңлә, 0.779583 гөл_чәчәк, 0.778158 яралган, 0.775153 татлы, 0.772041 бөре, 0.770986 матурлык, 0.757603 кызыктыр, 0.754347 сугар, 0.754283 аклык, 0.748151 туклыклы, 0.747967 ямьле;

***эт – хайван, алмагач** – 0.555482 үсемлек, 0.532633 агач_куак, 0.522246 чыршы_нарат, 0.520003 җиләк_җимеш_куак, 0.513412 декоратив_үсемлек, 0.504221 җимеш, 0.501679 имән_юкә, 0.492185 яфраклы_агач, 0.491171 чия_слива, 0.490028 җиләк_җимеш, 0.489917 үсенте, 0.48975 алмагач_чия, 0.489199 дару_үлән, 0.48613 агач_куаклык, 0.484642 агач_үсенте, 0.479097 нарат_чыршы, 0.478646 ылыслы, 0.478233 ылыслы_агач, 0.477539 куак_үсенте, 0.474722 үсенте_утырт, 0.474009 слива, 0.470146 тәҗрибәле_бакчачы, 0.46699 чәчәк_бөрелән, 0.466524 алмагач_груша, 0.462612 җимеш_агач_куак, 0.462331 каен, 0.460904 бакча_җиләк, 0.458814 каен_нарат, 0.458287 кәлшә, 0.457976 агач;*



4 нче график

эт – хайван, карга *– 0.529546 кош, 0.452673 җәнлек_кош_корт, 0.45135 хайван_кош_корт, 0.424303 кош_оя, 0.417403 күз_кар_чукы, 0.417002 киек_кош, 0.415236 кош_корт_хайван, 0.413669 җәнлек_җанвар, 0.408859 бөҗәк, 0.407776 тереклек_ия, 0.403236 җәнлек_кош, 0.40039 ерткыч_кош, 0.394989 дуңгыз_асрау, 0.394723 бүдәнә, 0.388843 кыргый_хайван, 0.387124 кош_җәнлек, 0.384505 чыпчык, 0.384216 җәнлек, 0.382977 бытбылдык, 0.380001 күчмә_кош, 0.379483 карр_карр, 0.379233 зарарлы_бөҗәк, 0.379104 ерткыч_хайван, 0.376883 кыргый_җәнлек, 0.376299 мәхлукат, 0.374626 каргыш, 0.371003 җәнлек_хайван, 0.368932 хайван_үсемлек, 0.368489 кош_корт, 0.366938 үрдәк_каз;*

җир_шар – планета, кояш *– 0.925748 йолдыз, 0.911985 якты, 0.909303 балкы, 0.897471 яктылык, 0.879299 болыт_капла, 0.878039 томанлы, 0.872683 болыт, 0.868168 нур_сип, 0.863357 балкыш, 0.862675 нур, 0.858728 яктыр, 0.857401 томан, 0.855835 сүн, 0.852218 ап_ачык, 0.850376 офык, 0.848987 кояш_нур, 0.848596 аяз, 0.844198 зәңгәр_күк, 0.843785 күк, 0.842158 сурән, 0.838735 сихри, 0.83776 кояш_балкы, 0.83621 нурлан, 0.826901 болытлы, 0.826159 яшәр, 0.825893 күзлә, 0.820815 балкытып, 0.818787 нурлы, 0.818328 алланып, 0.818247 җансыз;*

умарта_корт – бал, сыер – *0.560947 литр_сөт, 0.52204 сөт_сау, 0.512419 килограмм_сөт, 0.508482 кг_сөт_сау, 0.508447 майлылык, 0.503198 мең_килограмм_сөт, 0.49752 литр_сөт_сау, 0.496192 сөт, 0.482043 тонна_сөт, 0.479899 сау_сыер, 0.477969 сыер_сөт_сау, 0.474186 сыер_тәүлек, 0.474164 сөт_литр, 0.470264 көнлек_савым, 0.469839 тәүлеклек_савым, 0.468629 тана, 0.467971 килограмм_тәшкил, 0.466081 савым, 0.463965 артык_сөт_сау, 0.463271 сөт_майлылык, 0.460077 урта_килограмм_сөт, 0.45934 шикәр_ком, 0.459334 эремчек, 0.458588 сыер_сау, 0.458409 бозаула_тана, 0.457793 сыер_савым, 0.457119 майлылык_процент, 0.45479 сыер_көнлек_савым, 0.453182 сыер_сөт, 0.452489 тонна_сөт_сау.*

3. Ике сүз арасындагы мәгънәви якынлык күрсәткече. Әлеге функция югарыда каралган гамәлләрнең нигезендә ята. Ул ике сүзнең мәгънә ягыннан якынлыгын косинус охшашлыгы формуласы нигезендә исәпләп чыгара:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

Биредә **A** белән **B** – сүзләрне ачыклый торган векторлар, ә A_i белән B_i – әлеге векторларның компонентлары [Cosine similarity]. Мәсәлән, «*эт*» белән «*бүре*» сүзләре «*урындык*» белән «*сыер*» сүзләренә караганда бер-берсенә якынрак торалар: *0.543926 эт бүре 0.111563 урындык сыер.*

Ә «*тычкан*» сүзенең «*китап*», «*компьютер*» һәм «*күсе*» сүзләре белән мөнәсәбәтендә «*компьютер – тычкан*» пары «*китап – тычкан*» парына караганда якынрак, чөнки бу очракта «*тычкан*» сүзенең компьютер җиһазын белдерә торган мәгънәсе дә чагыла: *0.661968 күсе тычкан, 0.178382 компьютер тычкан, 0.079947 китап тычкан.*

Нәтиҗә. Word2vec моделе нигезендә әлеге эш кысаларында ясалган «Тезаурус» системасы, беренче чиратта, семантика һәм лексикография өлкәсендәге эзләнүләрдә файдалы булырга тиеш. Нейрон челтәрләр технологиясе кайбер очракларда шактый кечкенә күләмле корпус кысаларында да бай мәгълүмат бирә ала. Ләкин моның өчен күпсанлы көйләү параметрларын истә тотарга кирәк: тәрәзәнең киңлеге, итерацияләр саны, сүз векторларының зурлыгы һ. б. Гомумән алганда исә, барлык сүзләргә карата да яхшы нәтиҗәләргә ирешү өчен, берничә миллиард сүзне үз эченә алган югары сыйфатлы корпусның булуы зарур.

Әдәбият

Анализ тональности текста // Википедия. URL: https://ru.wikipedia.org/wiki/Анализ_тональности_текста (дата обращения: 05.03.2019).

Дистрибутивная семантика // Википедия. URL: https://ru.wikipedia.org/wiki/Дистрибутивная_семантика (дата обращения: 05.03.2019).

Татар матур әдәбияты корпусы. URL: http://litcorpus.antat.ru (дата обращения: 05.03.2019).

Татар теленең язма корпусы. URL: http://www.corpus.tatar/tt (дата обращения: 05.03.2019).

Cosine similarity // Wikipedia. URL: https://en.wikipedia.org/wiki/Cosine_similarity (дата обращения: 05.03.2019).

Natural language processing // Wikipedia. URL: https://en.wikipedia.org/wiki/Natural_language_processing (дата обращения: 05.03.2019).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems. Volume 2. Pages 3111–3119. https://arxiv.org/abs/1301.3781

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. ICLR Workshop, 2013. https://arxiv.org/abs/1310.4546

Word2vec. Google Code. URL: https://code.google.com/archive/p/word2vec/ (дата обращения: 05.03.2019).

Сәйхунов Мансур Рифкать улы,
филология фәннәре кандидаты, ТР ФА Г. Ибраһимов исемендәге Тел, әдәбият һәм сәнгать институтының гомуми лингвистика бүлеге өлкән фәнни хезмәткәре

Хөсәенова Альбина Марат кызы,
Иннополис университетының Машиналы өйрәнү һәм белем тасвирлау лабораториясе ассистенты

Хөсәенов Рөстәм Рафаэль улы,
GDC компаниясе инженеры