Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

# Интелигентни системи - проект

# Документација

## Кластерирање на огласи за работа - jobposts

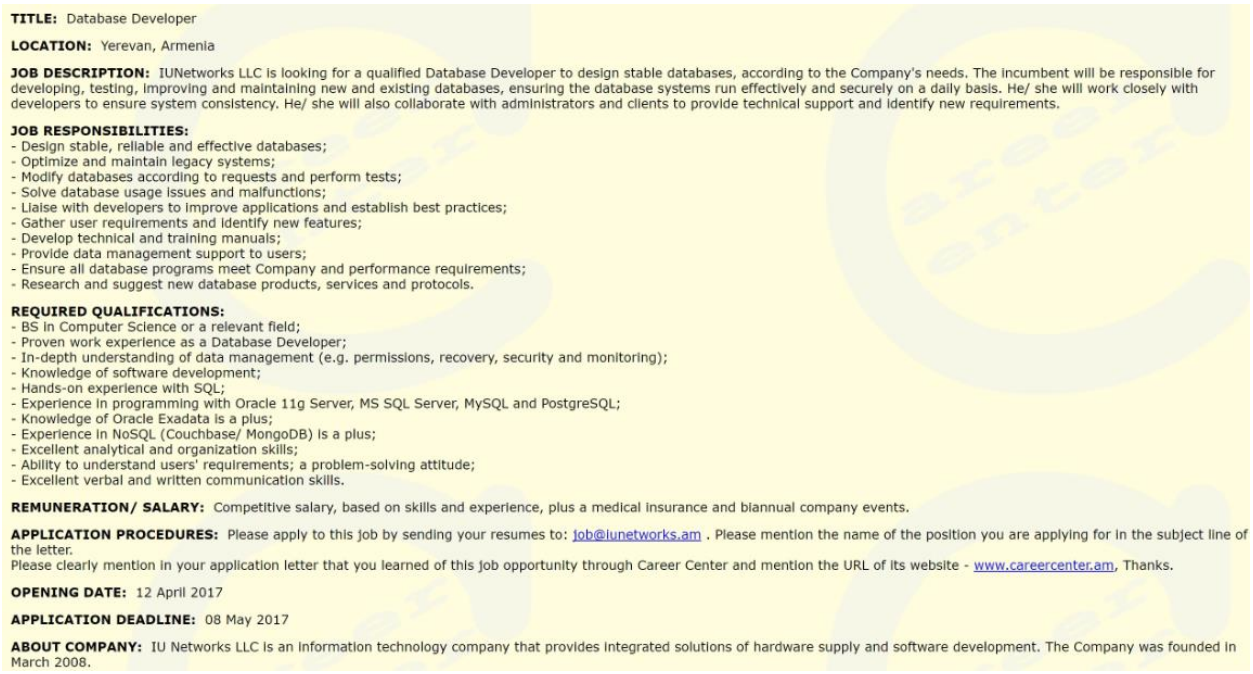

**Изработил:**

**Дарко Тосев 111224**

## Опис на проблемот

Целта на овој проект е поделбата на огласи за работа во секторите или областите во коишто припаѓаат, врз основа на нивната содржина.

Податочното множество коешто го имав на располагање е превземено од Kaggle[1]. Тоа се состои од 19000 објави за работа во текстуална форма, што биле поставени на Арменскиот портал за човечки ресурси – CareerCenter. Примерок за еден оглас е прикажан на сликата доле.



**TITLE:** Database Developer

**LOCATION:** Yerevan, Armenia

**JOB DESCRIPTION:** IUNetworks LLC is looking for a qualified Database Developer to design stable databases, according to the Company's needs. The incumbent will be responsible for developing, testing, improving and maintaining new and existing databases, ensuring the database systems run effectively and securely on a daily basis. He/ she will work closely with developers to ensure system consistency. He/ she will also collaborate with administrators and clients to provide technical support and identify new requirements.

**JOB RESPONSIBILITIES:**
- Design stable, reliable and effective databases;
- Optimize and maintain legacy systems;
- Modify databases according to requests and perform tests;
- Solve database usage issues and malfunctions;
- Liaise with developers to improve applications and establish best practices;
- Gather user requirements and identify new features;
- Develop technical and training manuals;
- Provide data management support to users;
- Ensure all database programs meet Company and performance requirements;
- Research and suggest new database products, services and protocols.

**REQUIRED QUALIFICATIONS:**
- BS in Computer Science or a relevant field;
- Proven work experience as a Database Developer;
- In-depth understanding of data management (e.g. permissions, recovery, security and monitoring);
- Knowledge of software development;
- Hands-on experience with SQL;
- Experience in programming with Oracle 11g Server, MS SQL Server, MySQL and PostgreSQL;
- Knowledge of Oracle Exadata is a plus;
- Experience in NoSQL (Couchbase/ MongoDB) is a plus;
- Excellent analytical and organization skills;
- Ability to understand users' requirements; a problem-solving attitude;
- Excellent verbal and written communication skills.

**REMUNERATION/ SALARY:** Competitive salary, based on skills and experience, plus a medical insurance and biannual company events.

**APPLICATION PROCEDURES:** Please apply to this job by sending your resumes to: job@iunetworks.am . Please mention the name of the position you are applying for in the subject line of the letter.
Please clearly mention in your application letter that you learned of this job opportunity through Career Center and mention the URL of its website - www.careercenter.am, Thanks.

**OPENING DATE:** 12 April 2017

**APPLICATION DEADLINE:** 08 May 2017

**ABOUT COMPANY:** IU Networks LLC is an information technology company that provides integrated solutions of hardware supply and software development. The Company was founded in March 2008.

Како што може да се забележи во огласот има неколку карактеристични полиња коишто започнуваат со карактеристични зборови, коишто главно се повторуваат во поголемиот дел на огласи. Во податочното множество со коешто работев, според тие зборови беше извршена едноставна поделба на оригиналниот текст во огласите во неколку делови[2]:

| Име на полето | Опис |
|---|---|
| **jobpost** | Оригиналниот текст на целосниот оглас |
| **date** | Дата на постирање на порталот |
| **Title** | Наслов на работното место |
| **Company** | Работодавач |
| **AnnouncementCode** | Код на објава(некој внатрешен код, што најчесто недостасува) |
| **Term** | Целосно, половично работно време итн. |
| **Eligibility** | Соодветни кандидати |

| | |
|---|---|
| **Audience** | Кој може да аплицира? |
| **StartDate** | Датум на започнување со работа |
| **Duration** | Времетраење на вработувањето |
| **Location** | Локација на работното место |
| **JobDescription** | Опис на работното место |
| **JobRequirment** | Одговорности и задачи на работното место |
| **RequiredQual** | Потребни квалификации на кандидатот |
| **Salary** | Плата |
| **ApplicationP** | Процедура за аплицирање |
| **OpeningDate** | Датум на објавување на огласот |
| **Deadline** | Краен рок-датум на огласот |
| **Notes** | Дополнителни забелешки |
| **AboutC** | Опис на компанијата |
| **Attach** | Додатоци |
| **Year** | Година на објавата на огласот(земена од date) |
| **Month** | Месец на објавата на огласот(земена од date) |
| **IT** | Булеан вредност за припадноста на работното место во ИТ секторот |

Сепак, секој оглас ја нема истата структура(различни работодавци, различно ги структурирале огласите, или пак како последица од соодветните разлики во секој сектор на работа) и поради тоа има голем дел на полиња што не се пополнети во различни огласи.

Променливата IT е добиена со едноставно пребарување на некои професии во ИТ секторот(пр: Java software developer, System administrator, Quality assurance engineer и сл.) во полето Title, односно имаме вредност true доколку насловот го содржи името на таа професија и вредност false доколку не го содржи. Оваа вредност не е добиена според содржината на огласот и не е извршена било каква обработка на текстот од огласот за таа цел.

Според ова, како и недостатокот на било каков друг атрибут што би можел да се смета за некаква класна лабела(Title е поле со приближно уникатни наслови за секој оглас) можам да заклучам дека немаме никакви лабели за огласите и не е можна некаква класификација на податоците. Како последица, со ова множество можеме да извршиме само некакво кластерирање и да погледнеме какво ќе биде нивното групирање според содржината, што го одредив како цел во овој проект.

Поради несовршеното делење на оригиналниот текст на огласот во овие полиња, решив да го користам целосниот текст на секој оглас што се наоѓа во првото поле jobpost, па понатамошната работа опишана во оваа документација се однесува само на обработка на ова поле од сите огласи во податочното множ(со исклучок во евалуацијата за којашто го користам и полето Title). Дополнително, користењето на целосните информации што се достапни за секој оглас, а не само на делови од него, би требало да допринесат за подобри резултати.

Не користам некакво репрезентативно множ. затоа што немам никакви информации за распределбата на огласите според нивниот сектор. Најмногу што би можел да направам е случаен избор на огласите за да креирам помало множ. но тоа не би бил репрезентативен примерок. Поради оваа причина го користам целосното множ. од 19000 огласи.

## Методологија

Како работна околина во која ги градам моделите и вршам обработка на податоците го користам програмскиот јазик Python, заедно со различните достапни пакети и библиотеки за машинско учење и обработка на текст. Специфично, ги користам следните пакети:

- nltk пакетот[3] за предпроцесирањето на текстот и токенизацијата.
- sklearn пакетот[4] за векторизација, алгоритмите за кластерирање, pca и мерките за евалуација.
- spherecluster пакетот[5] за K-средини со косинусно растојание.
- genism пакетот[6] за имплементација на Word2vec моделот.
- Други стандардни пакети(numpy, pandas, …).

Обработката на податоците се одвиваше во неколку чекори:

**Чекор 1 – Предпроцесирање на текстот, токенизација**

За секој оглас се прави следното: се тргаат интерпукциските знаци и бројките, големи букви се претвораат во мали, се тргаат непотребните зборови(stopwords) и се врши лемитизирање на преостанатите зборови. За идентификација на непотребните зборови ја

користев листата на stopwords за англискиот јазик дефинирани nltk пакетот, а за лематизација – WordNetLemmatizer достапен во истиот пакет.

**Чекор 2 – Векторизација**

Користев два методи за векторизација на огласите.

**Bag of words**

Првиот начин е класичниот модел bag of words, односно за секој оглас создаваме вектор на фреквенциите на зборовите во огласот. Како метод за пресметување на фреквенциите на зборовите ја користев нивната tf-idf(Term frequency – Inverse document frequency) вредност, којашто ја зима во предвид и нивната генерална честота во сите огласи. За потребите на овој чекор, го користев TfidfVectorizer во sklearn пакетот. Како број на атрибути се земени првите 300 зборови подредени според нивната фреквенција. Би можеле да тестираме и со поголем број на зборови(некој број што го одликува бројот на просечни уникатни зборови во целиот корпус, што е поголем од 300), но поради ограничувања со RAM меморија не бев во можност да испробам поголеми димензии за овој модел.

**Word2vec**

Вториот начин е користење на Word2vec модел[7]. Овој модел е имплементација на невронска мрежа, објавена од Google, којашто учи дистрибуирани репрезентации на зборови[8]. Доколку се користи големо податочно множество(десетици билиони зборови) моделот продуцира вектори на зборови со интересни карактеристики. Зборовите со слично значење се појавуваат во кластери и векторите на тесно поврзани зборови(како на пр. аналогии) можат да се добијат преку векторска математика.

За потребите на овој проект, јас го употребив готовиот истриниран Word2vec модел[9] од Google, којшто содржи 300-димензионални вектори за 3 милиони зборови и фрази. Сепак, поради технички ограничувања, ги користам само првите 60000 истренирани зборови.

Понатаму за секој оглас конструирам вектор којшто претставува просек на векторскиот збир од сите зборови во огласот за коишто имаме истреniран вектор во Word2vec моделот. Димензиите на истренираните зборови во Word2vec моделот е 300, па и резултантниот вектор за огласот ќе има 300 атрибути. На овој начин добиваме репрезентативен вектор за целиот оглас, што ќе биде многу попрецизен доколку се вклучени најголемиот број на зборови во него, што зависи од бројот на зборови на коишто е истреniран Word2vec моделот.

**Чекор 3 – Редукција на димензионалноста**

Редукција на димензионалноста по двата начина на векторизација правам поради две причини:

- Поради ретко пополнетата матрица по векторизацијата добиена во првиот метод, како и поради големата димензионалност на матрицата по векторизацијата во двата методи, техниките за кластерирање коишто ги користам во следниот, не би дале добри резултати ако не се надминат овие проблеми. Ова е поради тоа што техниките го користат Евклидовото растојание како мерка за сличност на два вектори. Кај К-средини кластерирањето извршено над матрицата добиена од првиот метод на векторизација, овие проблеми се надминуваат со користење на косинусното растојание како мерка за сличност(поради ретко пополнетата матрица, ефективно ќе имаме и помала димензионалност), но сепак поради втората причина се решив и тука да ја редуцирам димензионалноста.
- Технички ограничувања – немав доволно RAM меморија, како и потребното долго процесирачко време за извршување.

Поради претходно наведеното, ја користев имплементацијата на Principal Component Analysis во пакетот sklearn за да ја редуцирам димензионалноста до 10 атрибути(првите 10 компоненти).

**Чекор 4 – Кластерирање**

За кластерирање искористив неколку од техниките имплементирани во пакетот sklearn, како и имплементацијата за К-средини со косинусно растојание во пакетот spherecluster.

Беа извршени:

- К-средини кластерирање со Евклидово растојание и косинусно растојание, со 6 до 10 кластери.
- Агломеративно хиерархиско кластерирање, со Group average како мерка за блискост меѓу кластери и Евклидово растојание како мерка за сличност меѓу вектори, исто така со 6 до 10 кластери

Резултатите можат да се видат во табелата прикажана подолу.

## Евалуација и резултати

Поради природата на кластерирањето, евалуацијата на кластерите не е добро дефиниран и развиен дел од процесот на анализа со кластери. Во однос на мерки коишто не користат надворешна информација(класни лабели), во пакетот sklearn се достапни коефициентот на силуета и Калински-Харабаз индексот. Двете мерки ја користат кластерната кохесија и разделеност како индикација за добрината на кластерирањето.

Коефициентот на силуета дава вредност блиска до 1 за добро дефинирани и разделени кластери, додека вредност блиска до -1 за погрешно кластерирање. Вредности околу 0 индицираат на преклопувачки кластери.

Калински-Харабаз индексот, на сличен начин, дава висока вредност доколку кластерите се добро дефинирани и разделени.

Проблемот со овие мерки е тоа што добро ги оценуваат кластерите со глобуларен облик, а поради природата на проблемот(огласите најчесто и припаѓаат на повеќе сектори истовремено, или содржината на повеќе огласи е слична иако тие припаѓаат на различни сектори) најверојатно природната форма на кластерите не е таква.

Поради ова, користам друга мерка(крстена score во резултатите подоле) што би била добра индикација на квалитетот на кластерирањето, доколку претпоставката дека насловите(полето Title) ја рефлектираат содржината во огласот е точна.

Прво, за сите наслови на огласите во еден кластер се врши првичното предпроцесирање како што беше извршено на целиот оглас во првиот чекор. Потоа, се бараат најчестите (процесирани – лематизирани) зборови што се појавуваат во насловите на огласите од кластерот(за резултатите прикажани во табелите подоле, земени се првите 3 најчести збора). Како мерка за добрината на кластерот претставува процентот на наслови коишто содржат барем еден од најчестите зборови во насловите од овој кластер или барем еден од зборовите слични на овие најчести зборови. Сличните зборови се одредуваат според претходно споменатиот Word2vec модел користејќи ја most_similar функцијата од genism пакетот. Како мерка за добрината на целосното кластерирање се зема просекот на ваквите проценти од сите кластери. Поголема вредност значи дека темите во огласите(нивниот сектор) во соодветните кластери се добро поделени.

Во табелата подоле се издвоени најдобрите резултати од секоја техника на кластерирање со зелена боја.

| **Векторизација** | **Кластерирање** | **Метрика / Параметар** | **Силуета** | **Калински-Харабаз индекс** | **Скор** |
|---|---|---|---|---|---|
| Tf-idf | Агломеративно хиерархиско | n_clusters=6, Евклидово, Average | 0.214300 | 2259.396843 | 0.72 |
| | | n_clusters=7, Евклидово, Average | 0.198166 | 2194.196711 | 0.73 |
| | | n_clusters=8, Евклидово, Average | 0.195033 | 2115.177339 | 0.70 |
| | | n_clusters=9, Евклидово, Average | 0.221539 | 2502.961717 | 0.70 |
| | | n_clusters=10, Евклидово, Average | 0.211003 | 2234.680082 | 0.72 |
| Word2vec | | n_clusters=6, Евклидово, Average | 0.155070 | 13.781628 | 0.70 |

| | | | | | |
|---|---|---|---|---|---|
| | | n_clusters=7, Евклидово, Average | 0.140637 | 11.874206 | 0.79 |
| | | n_clusters=8, Евклидово, Average | 0.111392 | 16.105727 | 0.79 |
| | | n_clusters=9, Евклидово, Average | 0.080606 | 75.983398 | 0.74 |
| | | n_clusters=10, Евклидово, Average | 0.065197 | 68.304196 | 0.74 |
| Tf-idf | К-средини, Евклидово | n_clusters=6 | 0.257955 | 3848.888652 | 0.67 |
| | | n_clusters=7 | 0.269446 | 3742.464450 | 0.67 |
| | | n_clusters=8 | 0.263581 | 3674.036365 | 0.68 |
| | | n_clusters=9 | 0.275460 | 3765.724149 | 0.68 |
| | | n_clusters=10 | 0.274151 | 3803.962836 | 0.69 |
| Word2vec | | n_clusters=6 | 0.177819 | 3310.168706 | 0.58 |
| | | n_clusters=7 | 0.162001 | 3033.459899 | 0.60 |
| | | n_clusters=8 | 0.153943 | 2841.149403 | 0.58 |
| | | n_clusters=9 | 0.147565 | 2648.475764 | 0.58 |
| | | n_clusters=10 | 0.148280 | 2500.078280 | 0.61 |
| Tf-idf | К-средини, косинусно | n_clusters=6 | 0.228922 | 3631.172938 | 0.63 |
| | | n_clusters=7 | 0.231407 | 3535.086114 | 0.63 |
| | | n_clusters=8 | 0.243857 | 3506.979184 | 0.66 |
| | | n_clusters=9 | 0.254067 | 3570.116688 | 0.67 |
| | | n_clusters=10 | 0.253476 | 3613.475702 | 0.66 |
| Word2vec | | n_clusters=6 | 0.170087 | 3229.029184 | 0.58 |

| | | | | | |
|---|---|---|---|---|---|
| | | n_clusters=7 | 0.157475 | 2965.742473 | 0.59 |
| | | n_clusters=8 | 0.147048 | 2745.670165 | 0.58 |
| | | n_clusters=9 | 0.147951 | 2557.535618 | 0.61 |
| | | n_clusters=10 | 0.140197 | 2418.174119 | 0.63 |

Како што може да се види од резултатите, коефициентот на силуета главно дава вредности околу 0, што значи дека кластерите што ги добиваме се главно од неглобуларен облик(што се и очекуваше), поради што и не добиваме вредности блиски до 1.

Во однос на Калински-Харабаз индексот, сите вредности што ги добиваме се релативно високи(со исклучок на агломеративното хиерархиско градено врз Word2vec моделот). Најголемите вредности за оваа мерка се појавуваат кај кластерирањето добиено со К-средини и Евклидово растојание.

Една слабост на пресметаниот скор е неговата пристрасност кон кластерирања со неколку големи кластери и неколку мали кластери, односно просечната вредност ќе се зголеми значително доколку имаме неколку мали и чисти кластери во однос на оние кластерирања каде што просечната големина на секој кластер е релативно еднаква кај сите кластери. Ова би значело дека би имале голем скор иако кластерирањето можеби не е најдобро. Таков е случајот за хиерархиските кластерирања изградени врз основа на Word2vec моделот(доле е прикажан едно кластерирање визуелно). Ова е одразено и во вредностите на Калински-Харабаз индексот и ова е причината поради која единствено во тие кластерирања имаме мали вредности за овој индекс.

За да ги видиме на некој начин визуелно кластерирањата, подоле имам прикажано две од кластерирањата, каде што секој оглас е претставен на следниот начин: доколку насловот за тој оглас се не се броел како дел од позитивните примероци(односно не содржал некој од најчестите зборови во насловите од тој кластер или некој сличен на нив), тогаш огласот е прикажан со оригиналниот наслов, а доколку неговиот наслов се

броел како позитивен, тогаш тој е заменет со најчестиот или сличниот на него збор што го содржал.

На сликата подоле е прикажано кластерирањето во сина боја од табелата со резултати. Може да се види дека иако има добар скор, кластерите се многу лоши со само еден кластер кадешто се наоѓаат најголемиот дел од огласите и со неколку огласи поделени во неколку кластери(дополнително и со празен кластер). Поради оваа причина, скорот во вакви ретки случаеви не е добар показател и затоа треба да се гледа и просечната големина на секој кластер во кластерирањето.

| | European Reg | Senior Cre | C# Deskto | Logisticiar | Seller | Motion Gr | Shop Ope | Head of H | engineer | specialist- | specialist | specialist- | Assistant | specialist | manager | Lead Soft | Salesmen | User Inter | manager- | specia |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | information | | | | | | | | | | | | | | | | | | | |
| 3 | cleaner | | | | | | | | | | | | | | | | | | | |
| 4 | head | | | | | | | | | | | | | | | | | | | |
| 5 | paper | paper | | | | | | | | | | | | | | | | | | |
| 6 | methodology | | | | | | | | | | | | | | | | | | | |
| 7 | | | | | | | | | | | | | | | | | | | | |

За разлика од ова кластерирање, на следната слика е прикажано хиерархиското кластерирање со 7 кластери изградено врз tf-idf моделот кадешто големината на сите кластери е во просек слична, и во ова кластерирање веќе може да се забележи некакво издвојување на некои кластери како посебни сектори.

| | developer | developer | developer | engineer-arch | engineer | developer | developer | developer | developer | Leading Special | IT Department | developer | IT Administrat | developer | deve |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | IATP Distance | course | course | political | Forum at Upli | journalism | course | course | journalism | course | course | course | course | political | cour |
| 3 | specialist-con | assistant | Shop Cashier i | Travel Agent | specialist | assistant | Training Liaisor | Reseller/ Mer | Maternal & Chi | manager | HR Generalist | specialist-expe | Symposium Ev | assistant | mana |
| 4 | manager | Regional Retail | marketing | Translator/ Cc | marketing | specialist | manager | marketing | marketing | marketing | marketing | marketing | marketing | Fashion Prod | Tech |
| 5 | specialist | specialist | head-assistant | Kapan Branch | Credit Officer | Senior Broker | Brokerage Unit | specialist | head-assistant | specialist-expe | specialist | specialist | specialist | specialist | Mosl |
| 6 | Tax Adviser | Accounting/ Sc | Chief Specialis | financial-finar | accountant | Senior Assista | accountant | Accounting | Scheduling | Budget and Cos | accountant | accountant-auc | accountant-co | accountant | Payr |
| 7 | specialist | specialist-consi | project | project-constr | project | project | Administrative | project | specialist-expe | project | project | Translator / Int | Online Busine | work from h | proje |

Можеме да претпоставиме дека во кластерот 1 огласите се однесуваат за ИТ секторот, во кластерот 6 за сметководство и финансии, во кластерот 4 поголемиот број на огласи припаѓаат во секторот за маркетинг и менаџерство, кластерот 2 е поврзан со новинарството и така натаму за останатите кластери, кадешто темата не е толку очигледна.

## Заклучок

Генерално кластерирањата извршени со Word2vec моделот покажаа послаби резултати(особено хиерархиските кластерирања кадешто најголемиот број на огласи остануваат во 2-3 кластери), но тоа е најверојатно поради малиот број на зборови што ги

користам во овој модел и сигурен сум дека со користење на сите 3 милиони зборови понудени во моделот од Google, или пак дури и нов креиран модел истриниран над самото наше множ. комбинирано со други, би имале многу поголема прецизност во кластерирањето.

Кластерирањата извршени со tf-idf моделот покажаа слични резултати за сите кластерирања, со малку подобри резултати за хиерархиското кластерирање.

К-средини со косинусно растојание, изненадувачки има послаби резултати во однос на другите кластерирања, најверојатно поради редуцирањето на димензионалноста и очекувам овие резултати да се подобрат доколку не се користи тој чекор, или се тестира со повеќе вредности, за ова кластерирање.

К-средини со Евклидово растојание и хиерархиското кластерирање(изграден врз основа на tf-idf моделот) даваат слични резултати, со хиерархиското клас. малку подобри во однос на К-средини(поголем скор). Сепак, разликите се премногу мали за да издвоиме некое од кластерирањата како супериорно во однос на другите.

Можат да се подобрат голем дел од резултатите доколку би се испробале големиот број на опции и различни параметри што можат да се подесуваат во секој чекор од целото процесирање. На пример: во предпроцесирачкиот чекор може да се изврши дополнително стемирање, во bag of words моделот може да се подесува димензијата на векторите(различен број на зборови од 300) или да се креира друг модел на векторизација(bag of centroids со Word2vec моделот), во редукцијата на димензионалноста може да се тестира различно редуцирање или воопшто да не се редуцира и на крајот во последниот чекор, во пакетот sklearn постојат и мн. други техники на кластерирања, со голем број на подесувачки параметри коишто можеби би донеле подобри резултати за овој проблем.

*Конечно, дури и со ограниченият број на тестирани параметри и модели се добиваат добри резултати и можат да се издвојат повеќе кластери во поголемиот дел на кластерирања кадешто јасно може да се види темата(и секторот) во содржината на огласите во тој кластер.*

Понатаму, доколку би ги земале овие или подобри кластери како класни лабели за огласите, би можеле да вршиме различни класификации и би можеле да одговориме на барања како одредување на најдобриот сектор за работа според квалификациите на лицето од интерес или одредување на најбараните работни места во секој сектор и сл.

## Референци


https://www.kaggle.com/madhab/jobposts/home

https://www.slideshare.net/HabetMadoyan/it-skills-analysis-63686238

https://www.nltk.org/

http://scikit-learn.org/stable/index.html

https://pypi.org/project/spherecluster/0.1.2/

https://pypi.org/project/gensim/

https://code.google.com/archive/p/word2vec/

http://www.cs.toronto.edu/~bonner/courses/2014s/csc321/lectures/lec5.pdf

https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit