NATIONAL RESEARCH
UNIVERSITY

Система анализа текстов научных публикаций на основе методов машинного обучения

Научный Руководитель:

Крылов Владимир Владимирович

Автор:

Вороная Ксения

Цель работы

Цель работы – разработка системы по семантическому анализу набора документов, основанной на методах машинного обучения и NLP, для автоматической генерации summary (automatic summarization problem), определения количественных показателей семантической близости документов, и визуализации документов в числовом 3-х мерном пространстве.

Задачи

- Обзор и изучение литературы по интеллектуальному анализу текстов (text mining)
- Предварительная обработка текстов естественного языка
- Изучение и использование модели векторного представления слов в пространстве (word embedding), и модели глубокого нейронного обучения Word2Vec
- Нахождение семантически близких документов в наборе, используя TF-IDF матрицы и метрику cosine similarity
- Изучение и применение некоторых существующих алгоритмов построения summary
- Оценка качества полученных summary, на двух наборах документов из разных областей (schizophrenia area & big data standardization area)
- Поиск ключевых слов, фраз в тексте (key words)
- Наглядная визуализация документов в 3-х мерном пространстве для дальнейшего анализа

Актуальность темы

Основная идея summarization - состоит в том, чтобы найти репрезентативное подмножество данных (подмножество предложений текста), которое содержит информацию всего набора (полного текста).

Кто также занимается этой проблемой?

- Google Brain team – > Text summarization with TensorFlow (August, 2016) https://research.googleblog.com/2016/08/text-summarization-with-tensorflow.html
- Facebook – > “A Neural Attention Model for Abstractive Sentence Summarization by Facebook AI Research”, published Sep. 3, 2015. https://arxiv.org/abs/1509.00685
- IBM – > AlchemyAPI (acquired by IBM in 2015) – > “AlchemyLanguage is now Watson Natural Language Understanding”.

Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond by IBM Watson, published Aug 10, 2016. https://arxiv.org/abs/1602.06023

- Alex Rudnick, Ewan Klein and etc. – > Natural Language Toolkit (NLTK)
- Yahoo! – > Summly application, 2013

Предварительна обработка текстов

Токенизация - это разбиение текста на более мелкие части, токены. К токенам могут относиться как слова, так и биграммы или же n-граммы.

Удаление стоп-слов (шумовых слов) - слова, знаки, символы, которые самостоятельно не несут никакой смысловой нагрузки (предлоги, союзы и т.д.)

Лемматизация - процесс приведения словоформы к лемме, её нормальной (словарной) форме.

Стемминг - процесс нахождения основы слова для заданного исходного слова.

Word Embedding, Word2Vec модель

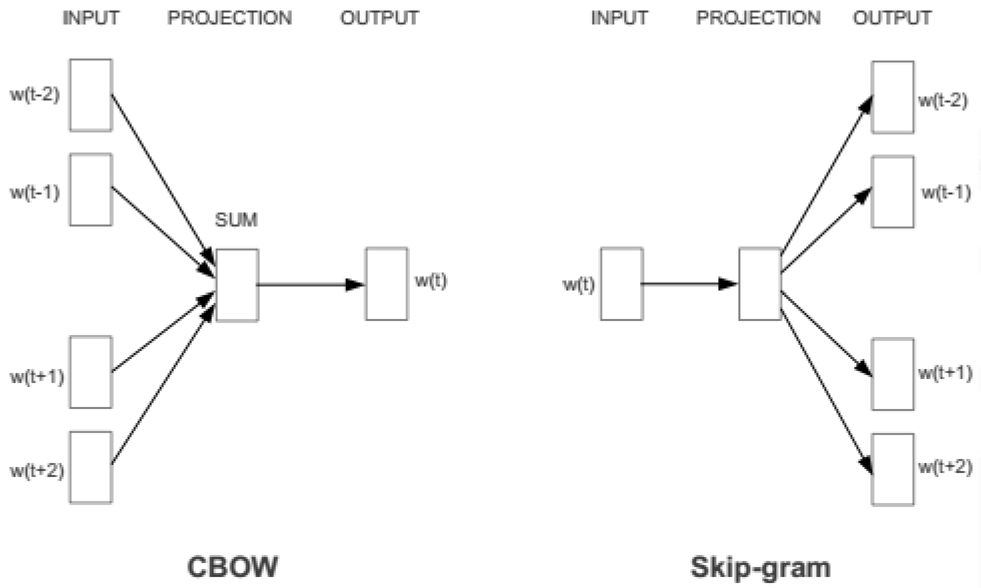
Word2vec модель – векторное представление слов.

$W: words \rightarrow R^n$ параметризованная функция, отображающая слова из некоторого естественного языка в векторы большой размерности.

В word2vec существуют два основных алгоритма обучения : CBOW (Continuous Bag of Words) и Skip-gram.

CBOW - предсказывает текущее слово, исходя из окружающего его контекста (learns a word by predicting its surrounding context).

Skip-gram - используем текущее слово, чтобы предугадывать окружающие его слова.



Word Embedding, Word2Vec модель

Vector space model (VSM) - представление коллекции документов векторами из одного общего для всей коллекции векторного пространства.

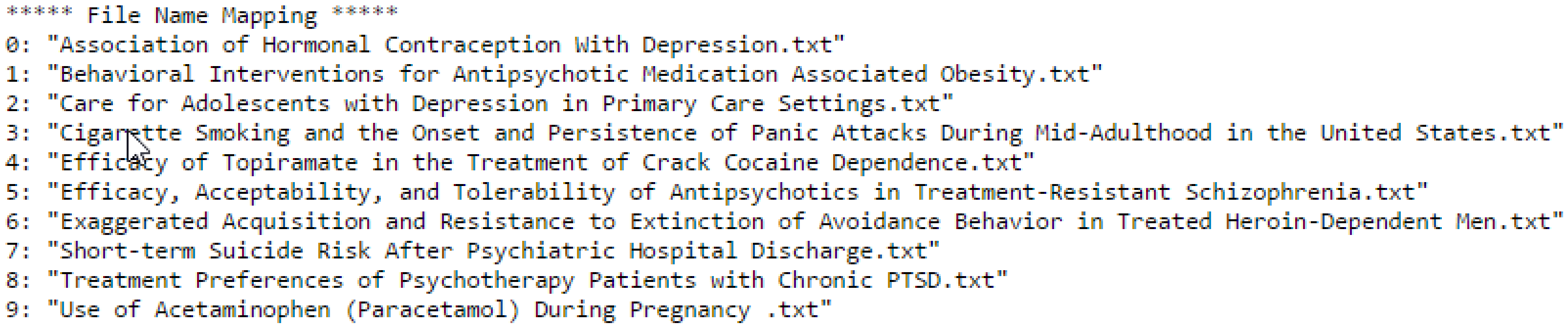
Документ в векторной модели рассматривается как неупорядоченное множество термов (слов, биграмм и т.д.).
Определяем вес (важность) терма в документе (можно tf или tf-idf или булевский вес).

$d_j = (w_{1j}, w_{2j}, \ldots, w_{nj})$ –> d_j принадлежит R^n, это вектор в VSM
d_j — векторное представление j-го документа,
w_{ij} — вес i-го терма в j-м документе,
n — общее количество различных термов во всех документах коллекции

Семантически близкие термы будут отображаться на близлежащие точки.

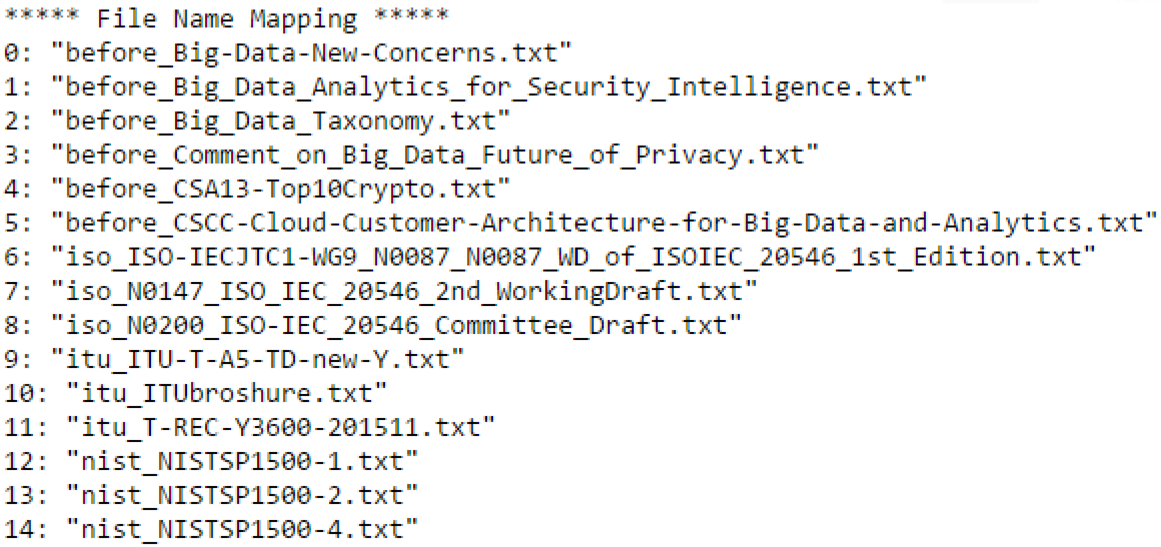
Datasets

Presented articles from the schizophrenia area.



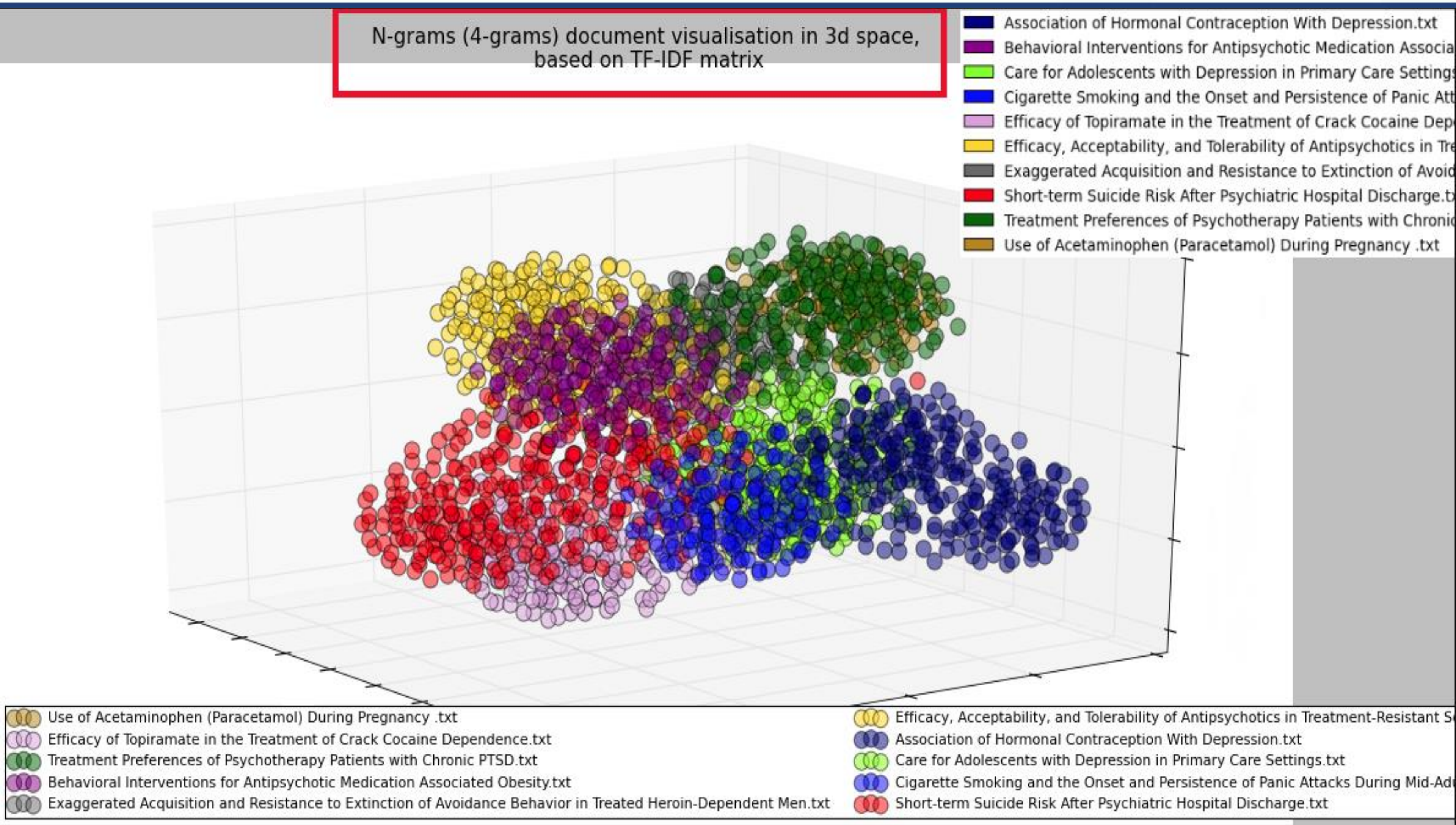
```
***** File Name Mapping *****
0: "Association of Hormonal Contraception With Depression.txt"
1: "Behavioral Interventions for Antipsychotic Medication Associated Obesity.txt"
2: "Care for Adolescents with Depression in Primary Care Settings.txt"
3: "Cigarette Smoking and the Onset and Persistence of Panic Attacks During Mid-Adulthood in the United States.txt"
4: "Efficacy of Topiramate in the Treatment of Crack Cocaine Dependence.txt"
5: "Efficacy, Acceptability, and Tolerability of Antipsychotics in Treatment-Resistant Schizophrenia.txt"
6: "Exaggerated Acquisition and Resistance to Extinction of Avoidance Behavior in Treated Heroin-Dependent Men.txt"
7: "Short-term Suicide Risk After Psychiatric Hospital Discharge.txt"
8: "Treatment Preferences of Psychotherapy Patients with Chronic PTSD.txt"
9: "Use of Acetaminophen (Paracetamol) During Pregnancy .txt"
```

Presented articles from the big data standardization area.

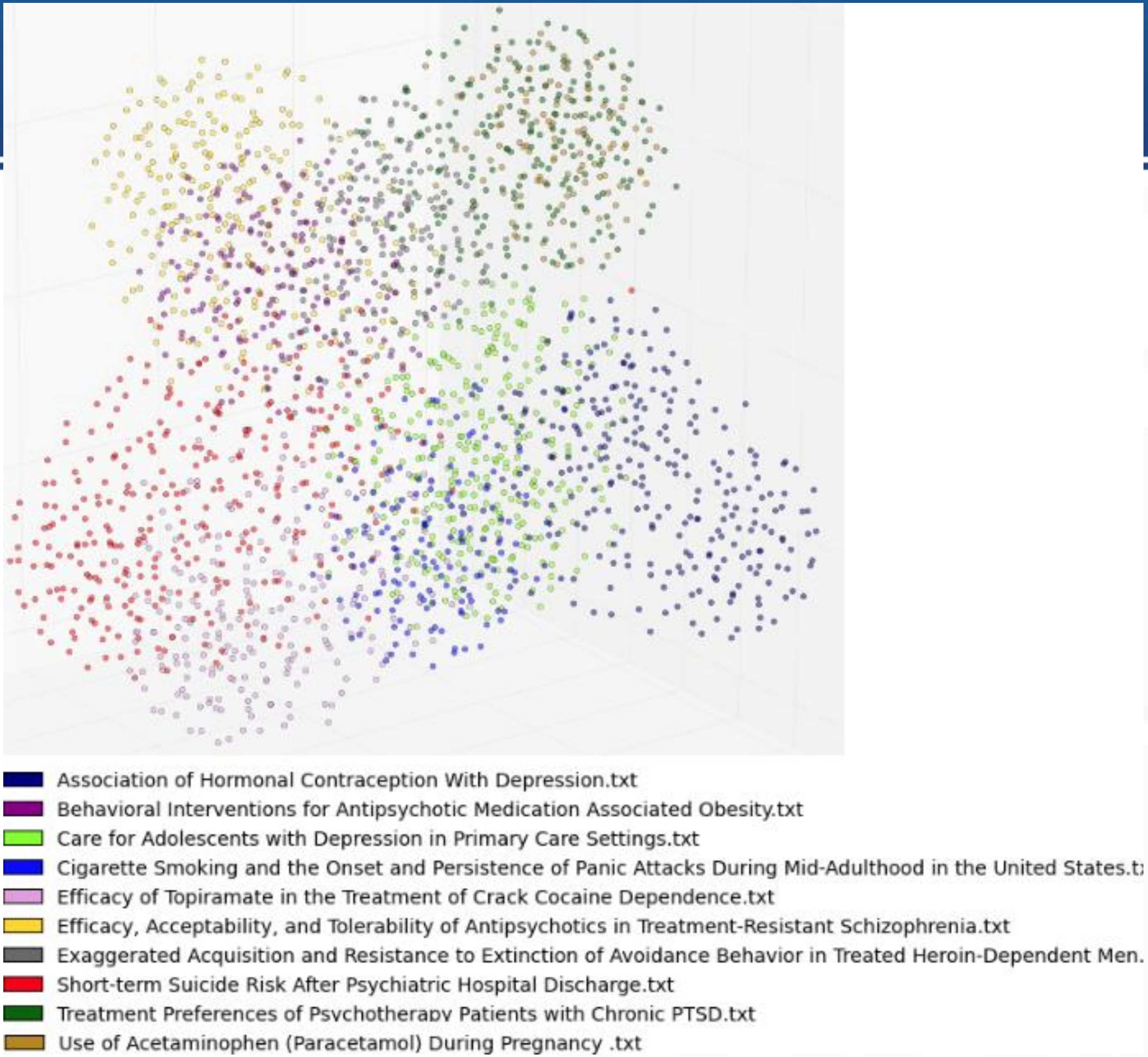


```
***** File Name Mapping *****
0: "before_Big-Data-New-Concerns.txt"
1: "before_Big_Data_Analytics_for_Security_Intelligence.txt"
2: "before_Big_Data_Taxonomy.txt"
3: "before_Comment_on_Big_Data_Future_of_Privacy.txt"
4: "before_CSA13-Top10Crypto.txt"
5: "before_CSCC-Cloud-Customer-Architecture-for-Big-Data-and-Analytics.txt"
6: "iso_ISO-IECJTC1-WG9_N0087_N0087_WD_of_ISOIEC_20546_1st_Edition.txt"
7: "iso_N0147_ISO_IEC_20546_2nd_WorkingDraft.txt"
8: "iso_N0200_ISO-IEC_20546_Committee_Draft.txt"
9: "itu_ITU-T-A5-TD-new-Y.txt"
10: "itu_ITUbroshure.txt"
11: "itu_T-REC-Y3600-201511.txt"
12: "nist_NISTSP1500-1.txt"
13: "nist_NISTSP1500-2.txt"
14: "nist_NISTSP1500-4.txt"
```

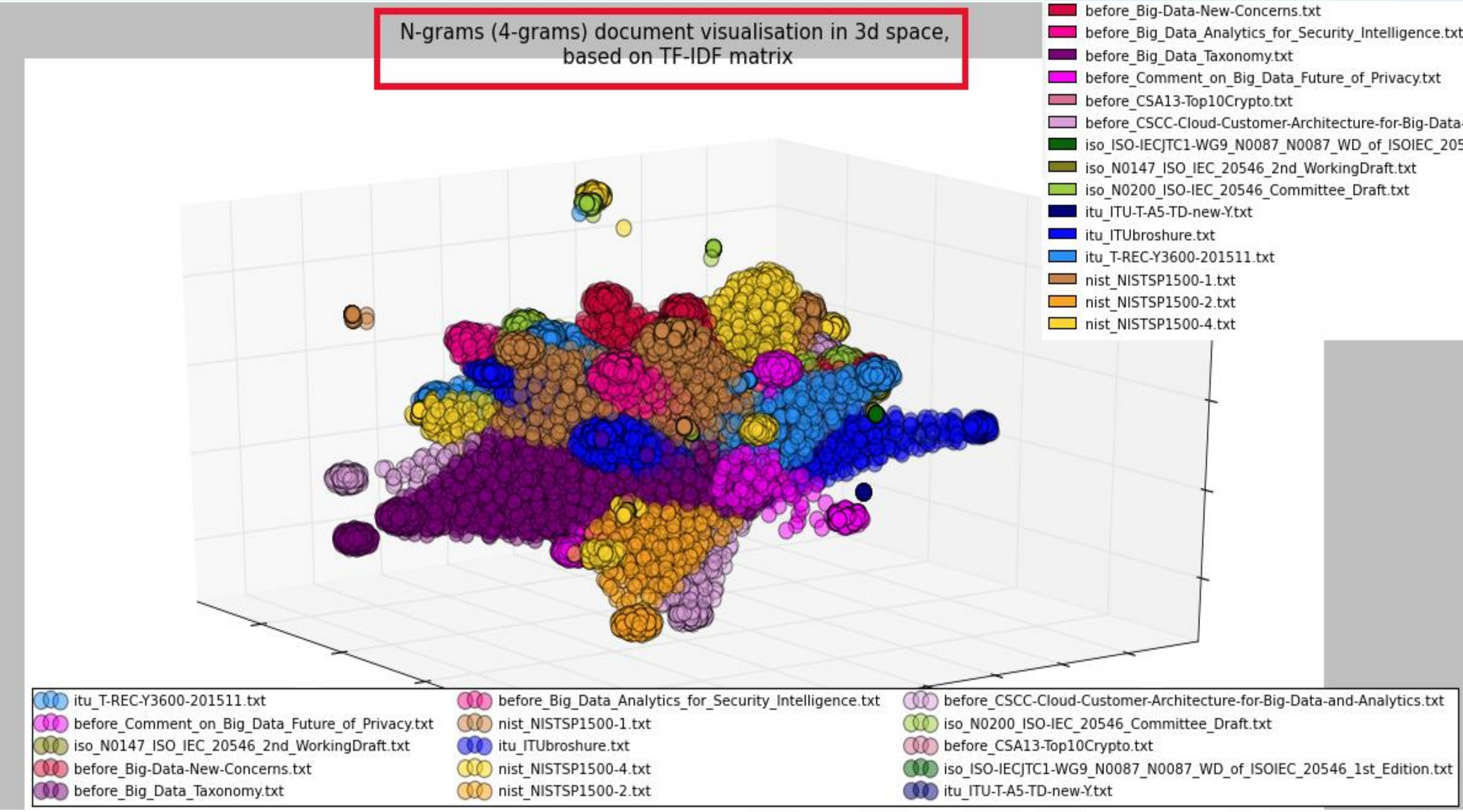
Dataset: texts from schizophrenia area



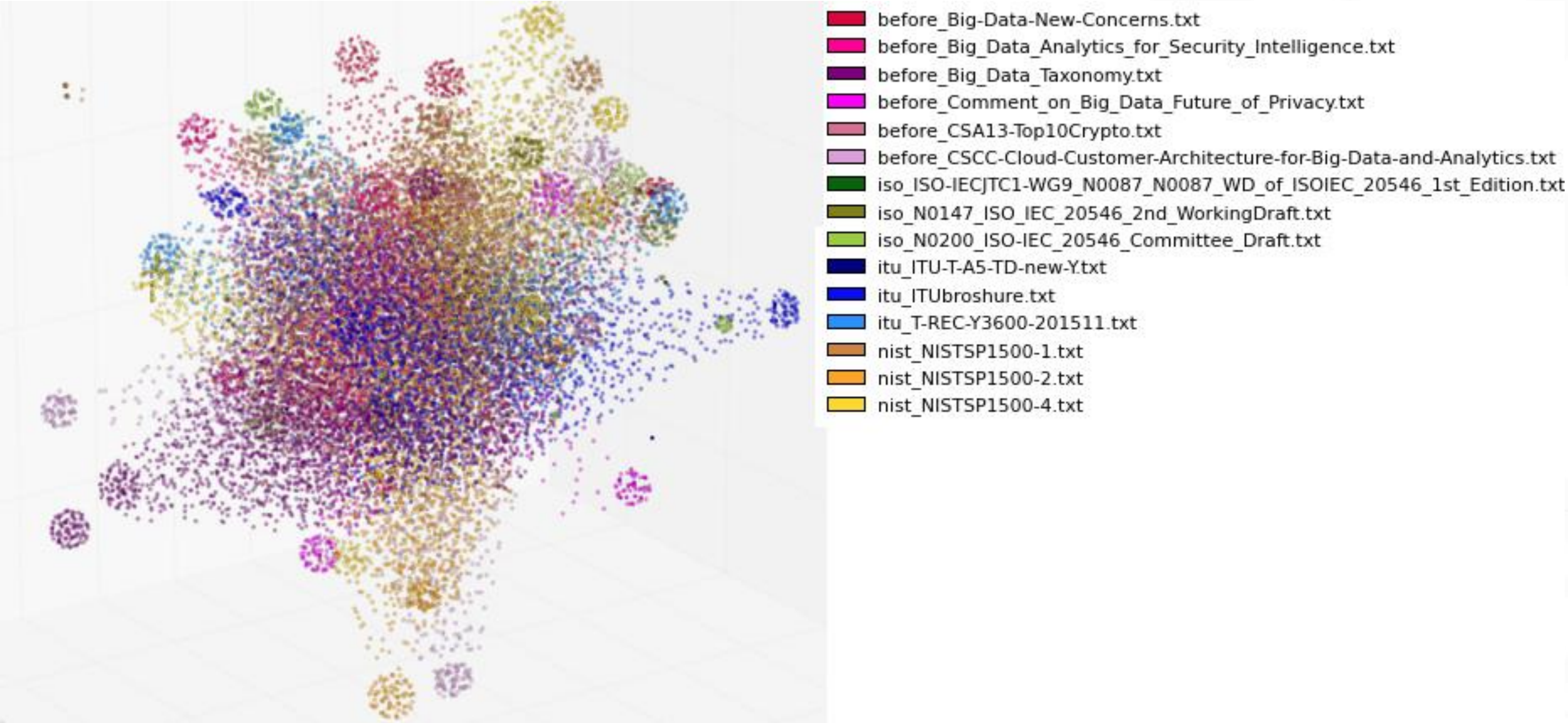
N-grams (4-grams) document visualisation in 3d space, based on TF-IDF matrix


Association of Hormonal Contraception With Depression.txt
Behavioral Interventions for Antipsychotic Medication Associated Obesity.txt
Care for Adolescents with Depression in Primary Care Settings.txt
Cigarette Smoking and the Onset and Persistence of Panic Attacks During Mid-Adulthood in the United States.t
Efficacy of Topiramate in the Treatment of Crack Cocaine Dependence.txt
Efficacy, Acceptability, and Tolerability of Antipsychotics in Treatment-Resistant Schizophrenia.txt
Exaggerated Acquisition and Resistance to Extinction of Avoidance Behavior in Treated Heroin-Dependent Men.
Short-term Suicide Risk After Psychiatric Hospital Discharge.txt
Treatment Preferences of Psychotherapy Patients with Chronic PTSD.txt
Use of Acetaminophen (Paracetamol) During Pregnancy .txt

Dataset: texts from big data standardization area



N-grams (4-grams) document visualisation in 3d space, based on TF-IDF matrix


before_Big-Data-New-Concerns.txt
before_Big_Data_Analytics_for_Security_Intelligence.txt
before_Big_Data_Taxonomy.txt
before_Comment_on_Big_Data_Future_of_Privacy.txt
before_CSA13-Top10Crypto.txt
before_CSCC-Cloud-Customer-Architecture-for-Big-Data-and-Analytics.txt
iso_ISO-IECJTC1-WG9_N0087_N0087_WD_of_ISOIEC_20546_1st_Edition.txt
iso_N0147_ISO_IEC_20546_2nd_WorkingDraft.txt
iso_N0200_ISO-IEC_20546_Committee_Draft.txt
itu_ITU-T-A5-TD-new-Y.txt
itu_ITUbroshure.txt
itu_T-REC-Y3600-201511.txt
nist_NISTSP1500-1.txt
nist_NISTSP1500-2.txt
nist_NISTSP1500-4.txt

Нахождение семантически близких документов в наборе, cosine similarity

Cosine Similarity (косинусное сходство) - это мера сходства между двумя векторами пространства, которая используется для измерения косинуса угла между ними.

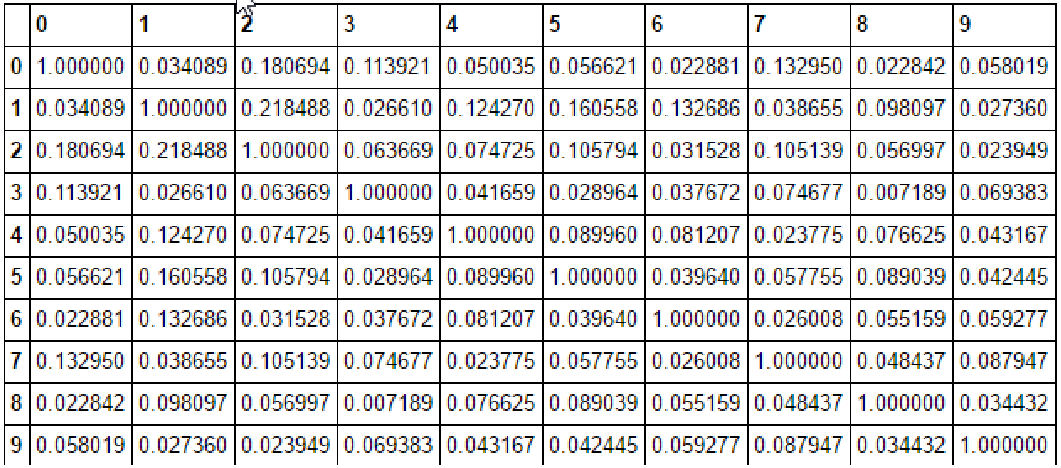
Если даны два вектора **A** и ***B***, то косинусное сходство, *cos(θ)*, может быть представлено используя скалярное произведение и норму:

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

Косинусное сходство эффективно в качестве оценочной меры, особенно для разреженных векторов, так как необходимо учитывать только ненулевые измерения.

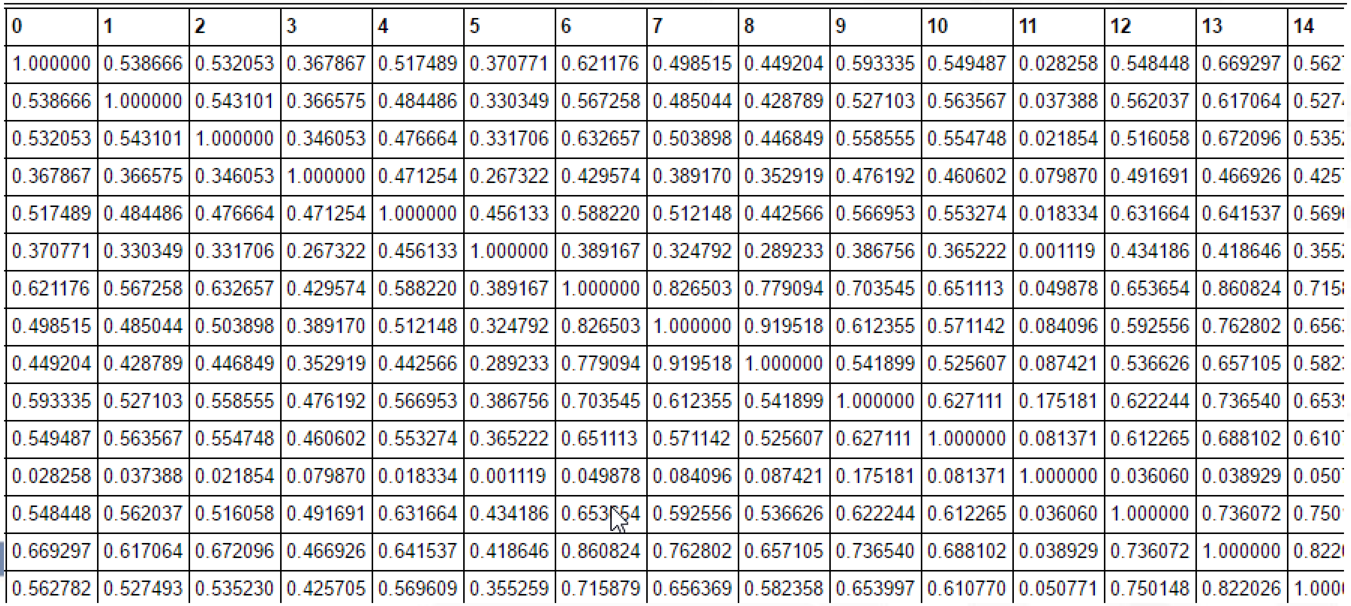
Поиск семантически близких документов, cosine similarity

Texts from schizophrenia area.



	0	1	2	3	4	5	6	7	8	9
0	1.000000	0.034089	0.180694	0.113921	0.050035	0.056621	0.022881	0.132950	0.022842	0.058019
1	0.034089	1.000000	0.218488	0.026610	0.124270	0.160558	0.132686	0.038655	0.098097	0.027360
2	0.180694	0.218488	1.000000	0.063669	0.074725	0.105794	0.031528	0.105139	0.056997	0.023949
3	0.113921	0.026610	0.063669	1.000000	0.041659	0.028964	0.037672	0.074677	0.007189	0.069383
4	0.050035	0.124270	0.074725	0.041659	1.000000	0.089960	0.081207	0.023775	0.076625	0.043167
5	0.056621	0.160558	0.105794	0.028964	0.089960	1.000000	0.039640	0.057755	0.089039	0.042445
6	0.022881	0.132686	0.031528	0.037672	0.081207	0.039640	1.000000	0.026008	0.055159	0.059277
7	0.132950	0.038655	0.105139	0.074677	0.023775	0.057755	0.026008	1.000000	0.048437	0.087947
8	0.022842	0.098097	0.056997	0.007189	0.076625	0.089039	0.055159	0.048437	1.000000	0.034432
9	0.058019	0.027360	0.023949	0.069383	0.043167	0.042445	0.059277	0.087947	0.034432	1.000000

Texts from big data standardization.



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1.000000	0.538666	0.532053	0.367867	0.517489	0.370771	0.621176	0.498515	0.449204	0.593335	0.549487	0.028258	0.548448	0.669297	0.562
0.538666	1.000000	0.543101	0.366575	0.484486	0.330349	0.567258	0.485044	0.428789	0.527103	0.563567	0.037388	0.562037	0.617064	0.527
0.532053	0.543101	1.000000	0.346053	0.476664	0.331706	0.632657	0.503898	0.446849	0.558555	0.554748	0.021854	0.516058	0.672096	0.535
0.367867	0.366575	0.346053	1.000000	0.471254	0.267322	0.429574	0.389170	0.352919	0.476192	0.460602	0.079870	0.491691	0.466926	0.425
0.517489	0.484486	0.476664	0.471254	1.000000	0.456133	0.588220	0.512148	0.442566	0.566953	0.553274	0.018334	0.631664	0.641537	0.569
0.370771	0.330349	0.331706	0.267322	0.456133	1.000000	0.389167	0.324792	0.289233	0.386756	0.365222	0.001119	0.434186	0.418646	0.355
0.621176	0.567258	0.632657	0.429574	0.588220	0.389167	1.000000	0.826503	0.779094	0.703545	0.651113	0.049878	0.653654	0.860824	0.715
0.498515	0.485044	0.503898	0.389170	0.512148	0.324792	0.826503	1.000000	0.919518	0.612355	0.571142	0.084096	0.592556	0.762802	0.656
0.449204	0.428789	0.446849	0.352919	0.442566	0.289233	0.779094	0.919518	1.000000	0.541899	0.525607	0.087421	0.536626	0.657105	0.582
0.593335	0.527103	0.558555	0.476192	0.566953	0.386756	0.703545	0.612355	0.541899	1.000000	0.627111	0.175181	0.622244	0.736540	0.653
0.549487	0.563567	0.554748	0.460602	0.553274	0.365222	0.651113	0.571142	0.525607	0.627111	1.000000	0.081371	0.612265	0.688102	0.610
0.028258	0.037388	0.021854	0.079870	0.018334	0.001119	0.049878	0.084096	0.087421	0.175181	0.081371	1.000000	0.036060	0.038929	0.050
0.548448	0.562037	0.516058	0.491691	0.631664	0.434186	0.653654	0.592556	0.536626	0.622244	0.612265	0.036060	1.000000	0.736072	0.750
0.669297	0.617064	0.672096	0.466926	0.641537	0.418646	0.860824	0.762802	0.657105	0.736540	0.688102	0.038929	0.736072	1.000000	0.822
0.562782	0.527493	0.535230	0.425705	0.569609	0.355259	0.715879	0.656369	0.582358	0.653997	0.610770	0.050771	0.750148	0.822026	1.000

Алгоритмы построения Summary

Два основных подхода к созданию summary:

- Извлечение (extraction)
- Обобщение (abstraction)

Извлекающие алгоритмы, которые анализируют текст статистически, а потом выбирают из него наиболее важные куски.

Обобщающие алгоритмы анализируют структуру текста, чтобы «понять», о чем он, а затем создают новый текст с основным содержанием.

В работе были рассмотрены следующие алгоритмы генерации summary:

1. **TextRank** (Rada Mihalcea and Paul Tarau, https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf)
2. **Latent Semantic Analysis, LSA** (Josef Steinberger http://www.kiv.zcu.cz/~jstein/publikace/isim2004.pdf)
3. **Kullback–Leibler divergence method for summarization** (A. Haghighi & Lucy Vanderwende http://www.aclweb.org/anthology/N09-1041)
4. **LexRank** (Güneş Erkan, Dragomir R. Radev https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume22/erkan04a-html/erkan04a.html)

TextRank

Идея:

1. построение графа на основе исходного текста на естественном языке
2. приближённое вычисление значения PageRank для построенного графа
3. применение полученных весов вершин для извлечения сведений из текста

В общем виде величина TextRank – это значение стационарного распределения случайного блуждания для каждой вершины $s \in V$ с учетом весов ребер.

$$\boldsymbol{TR(s_i) = (1 - \mathrm{d}) + \mathrm{d} \times \sum_{s_j \in In(s_i)} \frac{w_{ij}}{\sum_{s_k \in Out(s_j)} w_{jk}} \times TR(s_j)}$$

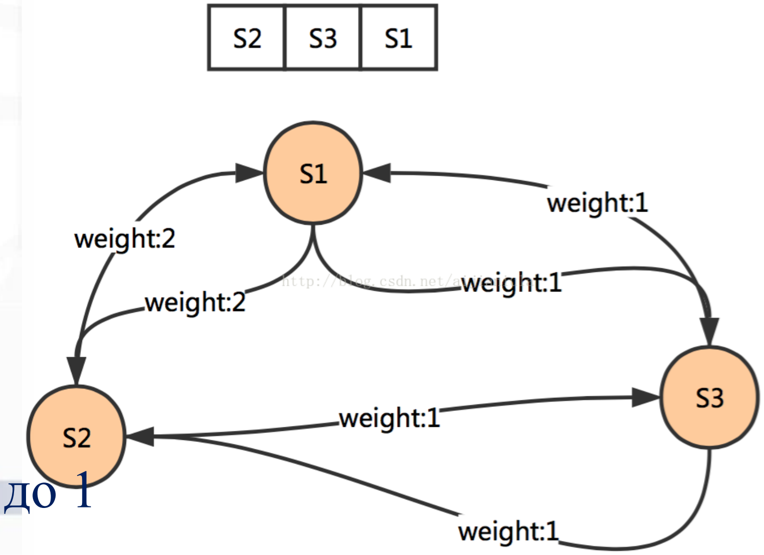
s – это предложение из текста (sentence)

In(s) – множество вершин входящих в s

Out(s) – множество вершин исходящих из s

w_{ij} - вес ребра (s_i, s_j)

d – фактор затухания, выбираем значение от 0 до 1



S2	S3	S1

Latent Semantic Analysis, LSA

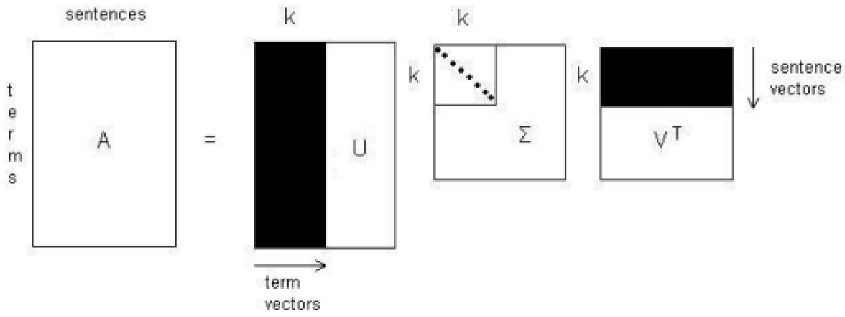
The latent semantic indexing, applied the singular value decomposition (SVD) to generic text summarization. The process starts with creation of a term by sentences matrix **A** = [A_1, A_2, …, A_n] with each column vector A_i, representing the weighted term-frequency vector of sentence *i* in the document under consideration. If there are a total of *m* terms and *n* sentences in the document, then we will have an *m* × *n* matrix **A** for the document. Since every word does not normally appear in each sentence, the matrix **A** is sparse.

Given an *m* × *n* matrix **A**, where without loss of generality *m* ≥ *n*, the SVD of **A** is defined as:

$$A = U\Sigma V^T ,$$

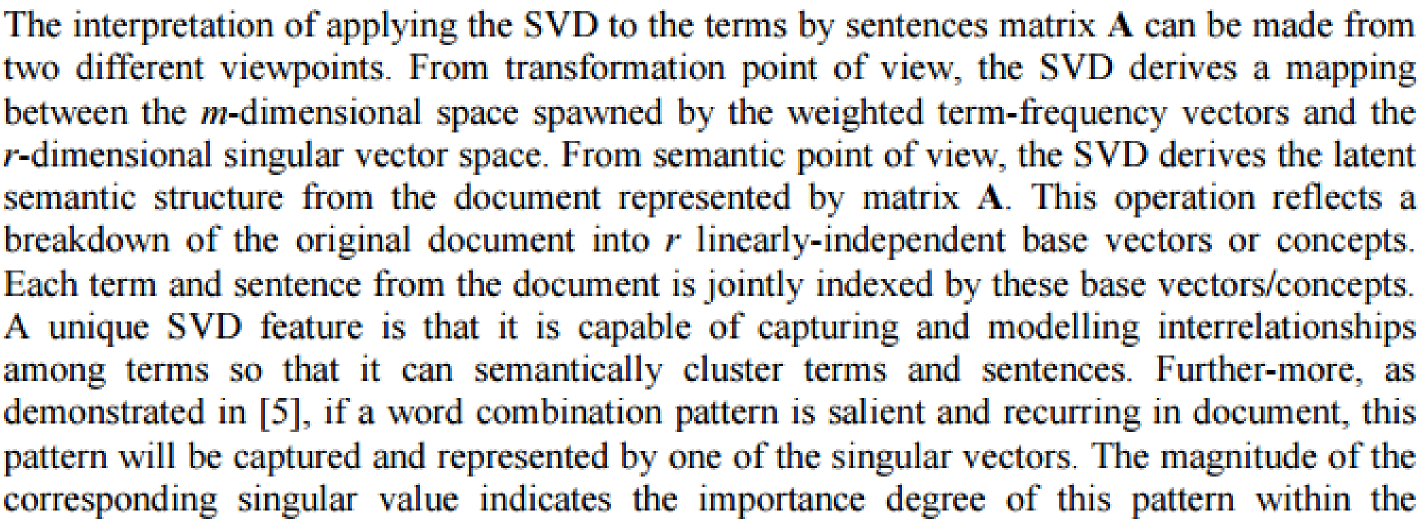
where **U** = [u_{ij}] is an *m* × *n* column-orthonormal matrix whose columns are called left singular vectors; **Σ** = diag(σ_1, σ_2, …, σ_n) is an *n* × *n* diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order, and **V** = [v_{ij}] is an *n* × *n* orthonormal matrix, whose columns are called right singular vectors (see figure 1). If rank(**A**) = *r*, then (see [5]) **Σ** satisfies:

$$\sigma_1 \geq \sigma_2 ... \geq \sigma_r > \sigma_{r+1} = ... = \sigma_n = 0 .$$

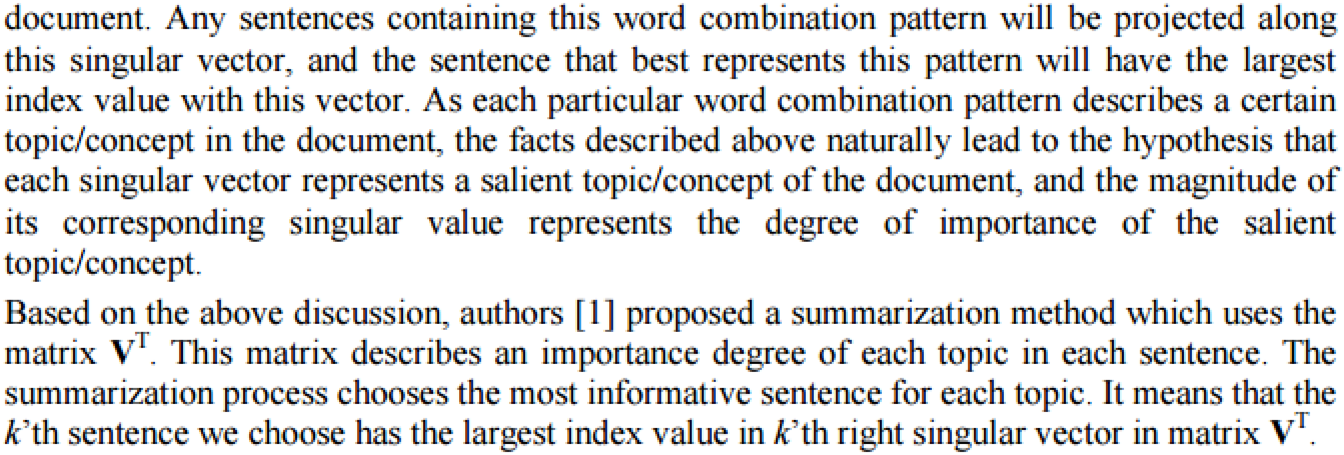


Singular Value Decomposition

Latent Semantic Analysis, LSA



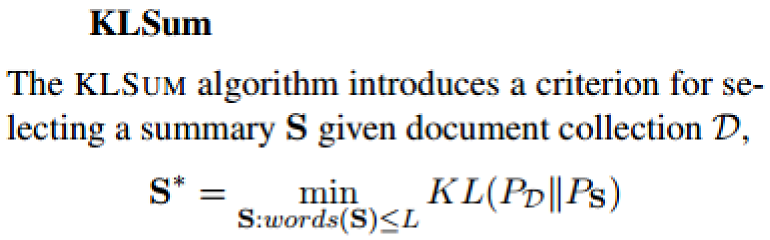
The interpretation of applying the SVD to the terms by sentences matrix **A** can be made from two different viewpoints. From transformation point of view, the SVD derives a mapping between the *m*-dimensional space spawned by the weighted term-frequency vectors and the *r*-dimensional singular vector space. From semantic point of view, the SVD derives the latent semantic structure from the document represented by matrix **A**. This operation reflects a breakdown of the original document into *r* linearly-independent base vectors or concepts. Each term and sentence from the document is jointly indexed by these base vectors/concepts. A unique SVD feature is that it is capable of capturing and modelling interrelationships among terms so that it can semantically cluster terms and sentences. Further-more, as demonstrated in [5], if a word combination pattern is salient and recurring in document, this pattern will be captured and represented by one of the singular vectors. The magnitude of the corresponding singular value indicates the importance degree of this pattern within the



document. Any sentences containing this word combination pattern will be projected along this singular vector, and the sentence that best represents this pattern will have the largest index value with this vector. As each particular word combination pattern describes a certain topic/concept in the document, the facts described above naturally lead to the hypothesis that each singular vector represents a salient topic/concept of the document, and the magnitude of its corresponding singular value represents the degree of importance of the salient topic/concept.

Based on the above discussion, authors [1] proposed a summarization method which uses the matrix $\mathbf{V}^T$. This matrix describes an importance degree of each topic in each sentence. The summarization process chooses the most informative sentence for each topic. It means that the *k*'th sentence we choose has the largest index value in *k*'th right singular vector in matrix $\mathbf{V}^T$.

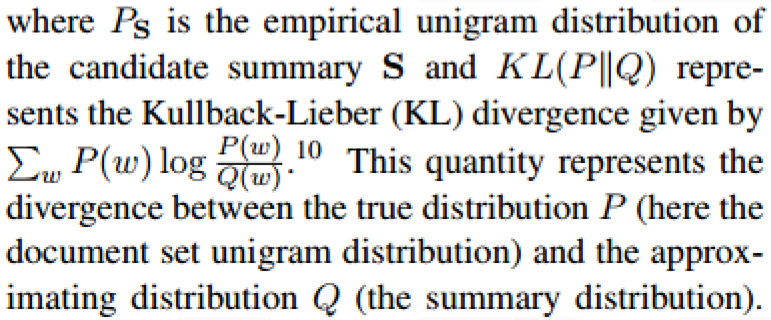
Kullback–Leibler divergence method for summarization



KLSum

The KLSum algorithm introduces a criterion for selecting a summary **S** given document collection $\mathcal{D}$,

$$\mathbf{S}^* = \min_{\mathbf{S}:words(\mathbf{S})\leq L} KL(P_{\mathcal{D}}\|P_{\mathbf{S}})$$



where $P_{\mathbf{S}}$ is the empirical unigram distribution of the candidate summary **S** and $KL(P\|Q)$ represents the Kullback-Lieber (KL) divergence given by $\sum_w P(w)\log\frac{P(w)}{Q(w)}$.[10] This quantity represents the divergence between the true distribution P (here the document set unigram distribution) and the approximating distribution Q (the summary distribution).

Kullback–Leibler divergence - это неотрицательнозначный функционал, являющийся несимметричной мерой удаленности друг от друга двух вероятностных распределений. Обычно одно из сравниваемых распределений — это «истинное» или постулируемое априори распределение (распределение P), второе — предполагаемое (проверяемое), являющееся приближением первого (распределение Q). Значение дивергенции является безразмерной величиной. Данная мера расстояния в теории информации может интерпретироваться как величина потерь информации при замене истинного распределения P на Q.

LexRank

Идея:

1. построение графа на основе исходного текста на естественном языке
2. приближённое вычисление значения PageRank для построенного графа
3. применение полученных весов вершин для извлечения сведений из текста

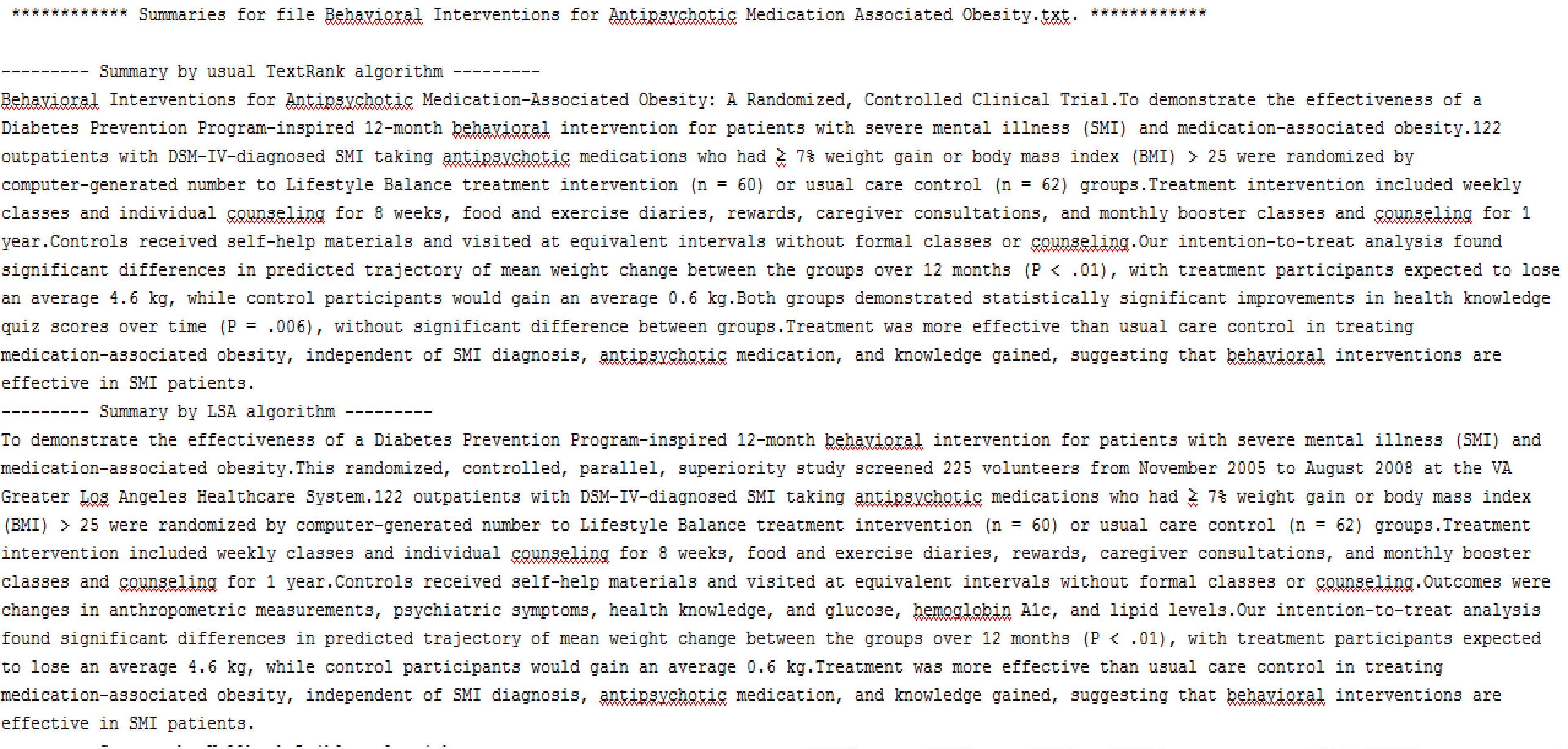
Отличие:

“LexRank based on the concept of eigenvector centrality in a graph representation of sentences. In this model, a connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences.”

LexRank uses cosine similarity of TF-IDF vectors.

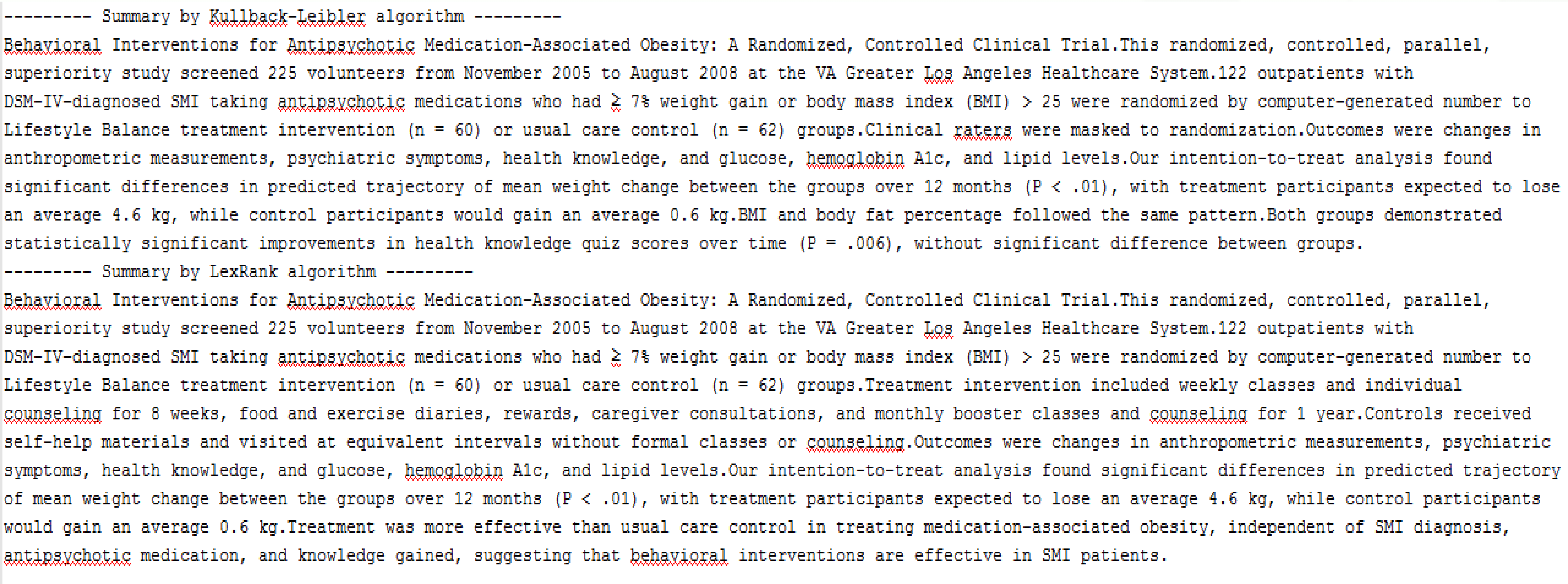
TextRank использует Random Walks, в то время как LexRank использует собственные вектора.

Примеры сгенерированного Summary для одного и того же документа, 4 разными алгоритмами



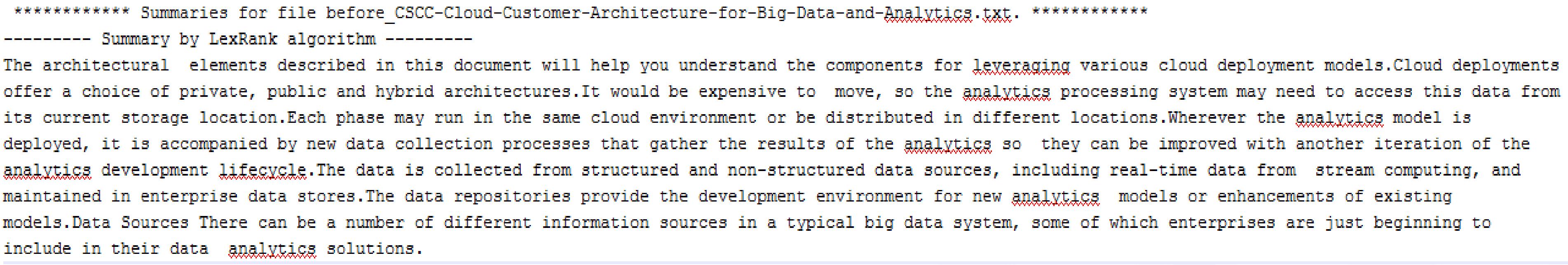
```
 ************ Summaries for file Behavioral Interventions for Antipsychotic Medication Associated Obesity.txt. ************

--------- Summary by usual TextRank algorithm ---------
Behavioral Interventions for Antipsychotic Medication-Associated Obesity: A Randomized, Controlled Clinical Trial.To demonstrate the effectiveness of a
Diabetes Prevention Program-inspired 12-month behavioral intervention for patients with severe mental illness (SMI) and medication-associated obesity.122
outpatients with DSM-IV-diagnosed SMI taking antipsychotic medications who had ≥ 7% weight gain or body mass index (BMI) > 25 were randomized by
computer-generated number to Lifestyle Balance treatment intervention (n = 60) or usual care control (n = 62) groups.Treatment intervention included weekly
classes and individual counseling for 8 weeks, food and exercise diaries, rewards, caregiver consultations, and monthly booster classes and counseling for 1
year.Controls received self-help materials and visited at equivalent intervals without formal classes or counseling.Our intention-to-treat analysis found
significant differences in predicted trajectory of mean weight change between the groups over 12 months (P < .01), with treatment participants expected to lose
an average 4.6 kg, while control participants would gain an average 0.6 kg.Both groups demonstrated statistically significant improvements in health knowledge
quiz scores over time (P = .006), without significant difference between groups.Treatment was more effective than usual care control in treating
medication-associated obesity, independent of SMI diagnosis, antipsychotic medication, and knowledge gained, suggesting that behavioral interventions are
effective in SMI patients.
--------- Summary by LSA algorithm ---------
To demonstrate the effectiveness of a Diabetes Prevention Program-inspired 12-month behavioral intervention for patients with severe mental illness (SMI) and
medication-associated obesity.This randomized, controlled, parallel, superiority study screened 225 volunteers from November 2005 to August 2008 at the VA
Greater Los Angeles Healthcare System.122 outpatients with DSM-IV-diagnosed SMI taking antipsychotic medications who had ≥ 7% weight gain or body mass index
(BMI) > 25 were randomized by computer-generated number to Lifestyle Balance treatment intervention (n = 60) or usual care control (n = 62) groups.Treatment
intervention included weekly classes and individual counseling for 8 weeks, food and exercise diaries, rewards, caregiver consultations, and monthly booster
classes and counseling for 1 year.Controls received self-help materials and visited at equivalent intervals without formal classes or counseling.Outcomes were
changes in anthropometric measurements, psychiatric symptoms, health knowledge, and glucose, hemoglobin A1c, and lipid levels.Our intention-to-treat analysis
found significant differences in predicted trajectory of mean weight change between the groups over 12 months (P < .01), with treatment participants expected
to lose an average 4.6 kg, while control participants would gain an average 0.6 kg.Treatment was more effective than usual care control in treating
medication-associated obesity, independent of SMI diagnosis, antipsychotic medication, and knowledge gained, suggesting that behavioral interventions are
effective in SMI patients.
```

Примеры сгенерированного Summary для одного и того же документа, 4 разными алгоритмами

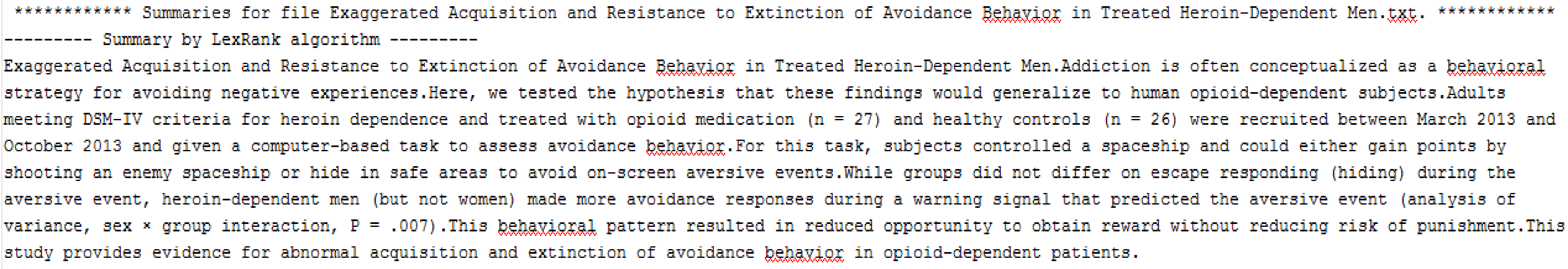


```
--------- Summary by Kullback-Leibler algorithm ---------
Behavioral Interventions for Antipsychotic Medication-Associated Obesity: A Randomized, Controlled Clinical Trial.This randomized, controlled, parallel,
superiority study screened 225 volunteers from November 2005 to August 2008 at the VA Greater Los Angeles Healthcare System.122 outpatients with
DSM-IV-diagnosed SMI taking antipsychotic medications who had ≥ 7% weight gain or body mass index (BMI) > 25 were randomized by computer-generated number to
Lifestyle Balance treatment intervention (n = 60) or usual care control (n = 62) groups.Clinical raters were masked to randomization.Outcomes were changes in
anthropometric measurements, psychiatric symptoms, health knowledge, and glucose, hemoglobin A1c, and lipid levels.Our intention-to-treat analysis found
significant differences in predicted trajectory of mean weight change between the groups over 12 months (P < .01), with treatment participants expected to lose
an average 4.6 kg, while control participants would gain an average 0.6 kg.BMI and body fat percentage followed the same pattern.Both groups demonstrated
statistically significant improvements in health knowledge quiz scores over time (P = .006), without significant difference between groups.
--------- Summary by LexRank algorithm ---------
Behavioral Interventions for Antipsychotic Medication-Associated Obesity: A Randomized, Controlled Clinical Trial.This randomized, controlled, parallel,
superiority study screened 225 volunteers from November 2005 to August 2008 at the VA Greater Los Angeles Healthcare System.122 outpatients with
DSM-IV-diagnosed SMI taking antipsychotic medications who had ≥ 7% weight gain or body mass index (BMI) > 25 were randomized by computer-generated number to
Lifestyle Balance treatment intervention (n = 60) or usual care control (n = 62) groups.Treatment intervention included weekly classes and individual
counseling for 8 weeks, food and exercise diaries, rewards, caregiver consultations, and monthly booster classes and counseling for 1 year.Controls received
self-help materials and visited at equivalent intervals without formal classes or counseling.Outcomes were changes in anthropometric measurements, psychiatric
symptoms, health knowledge, and glucose, hemoglobin A1c, and lipid levels.Our intention-to-treat analysis found significant differences in predicted trajectory
of mean weight change between the groups over 12 months (P < .01), with treatment participants expected to lose an average 4.6 kg, while control participants
would gain an average 0.6 kg.Treatment was more effective than usual care control in treating medication-associated obesity, independent of SMI diagnosis,
antipsychotic medication, and knowledge gained, suggesting that behavioral interventions are effective in SMI patients.
```

Примеры сгенерированных LexRank алгоритом Summaries, для документов из разных тем



```
 ************ Summaries for file before_CSCC-Cloud-Customer-Architecture-for-Big-Data-and-Analytics.txt. ************
--------- Summary by LexRank algorithm ---------
The architectural  elements described in this document will help you understand the components for leveraging various cloud deployment models.Cloud deployments
offer a choice of private, public and hybrid architectures.It would be expensive to  move, so the analytics processing system may need to access this data from
its current storage location.Each phase may run in the same cloud environment or be distributed in different locations.Wherever the analytics model is
deployed, it is accompanied by new data collection processes that gather the results of the analytics so  they can be improved with another iteration of the
analytics development lifecycle.The data is collected from structured and non-structured data sources, including real-time data from  stream computing, and
maintained in enterprise data stores.The data repositories provide the development environment for new analytics  models or enhancements of existing
models.Data Sources There can be a number of different information sources in a typical big data system, some of which enterprises are just beginning to
include in their data  analytics solutions.
```



```
 ************ Summaries for file Exaggerated Acquisition and Resistance to Extinction of Avoidance Behavior in Treated Heroin-Dependent Men.txt. ************
--------- Summary by LexRank algorithm ---------
Exaggerated Acquisition and Resistance to Extinction of Avoidance Behavior in Treated Heroin-Dependent Men.Addiction is often conceptualized as a behavioral
strategy for avoiding negative experiences.Here, we tested the hypothesis that these findings would generalize to human opioid-dependent subjects.Adults
meeting DSM-IV criteria for heroin dependence and treated with opioid medication (n = 27) and healthy controls (n = 26) were recruited between March 2013 and
October 2013 and given a computer-based task to assess avoidance behavior.For this task, subjects controlled a spaceship and could either gain points by
shooting an enemy spaceship or hide in safe areas to avoid on-screen aversive events.While groups did not differ on escape responding (hiding) during the
aversive event, heroin-dependent men (but not women) made more avoidance responses during a warning signal that predicted the aversive event (analysis of
variance, sex × group interaction, P = .007).This behavioral pattern resulted in reduced opportunity to obtain reward without reducing risk of punishment.This
study provides evidence for abnormal acquisition and extinction of avoidance behavior in opioid-dependent patients.
```

Оценка качества полученных Summary, ROUGE metric

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) – некий набор метрик, используемый для оценки качества автоматической summarization и машинного перевода, широко используется в NLP.

Author: **Chin-Yew Lin**,
Information Sciences Institute, University of Southern California
(http://www.aclweb.org/anthology/W04-1013)

“ It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans.”

Оценка качества полученных Summary, ROUGE metric

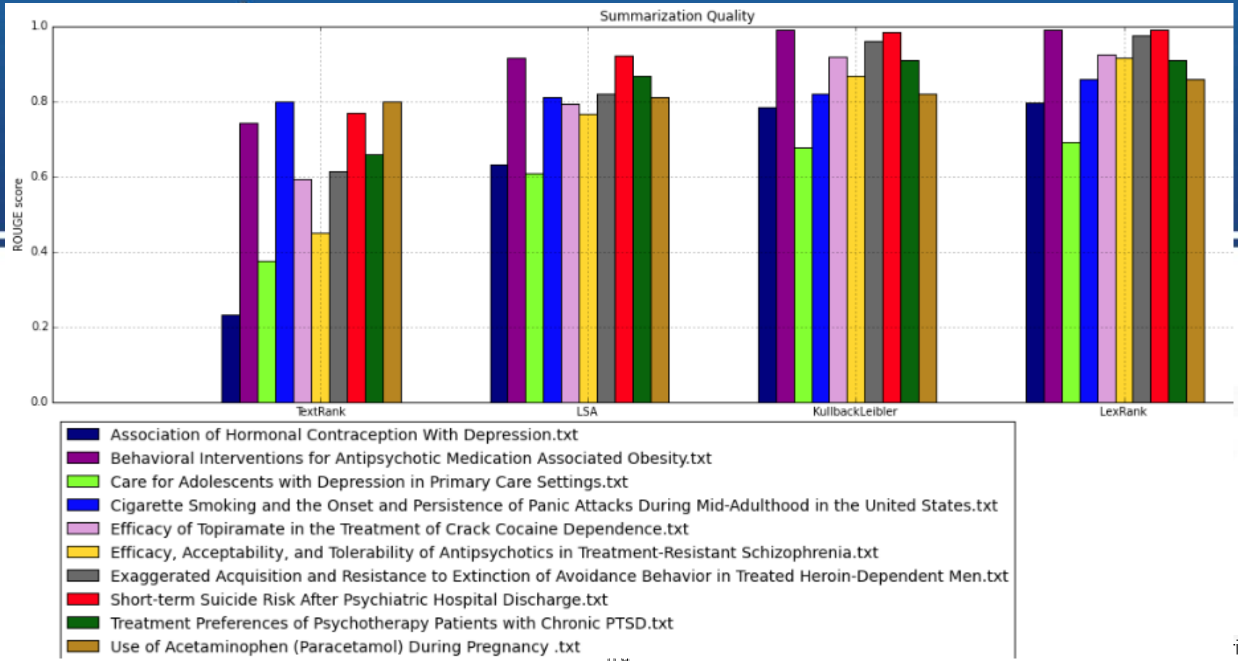
ROUGE-N (N-gram Co-Occurrence Statistics) - is an n-gram recall between a candidate summary and a set of reference summaries.

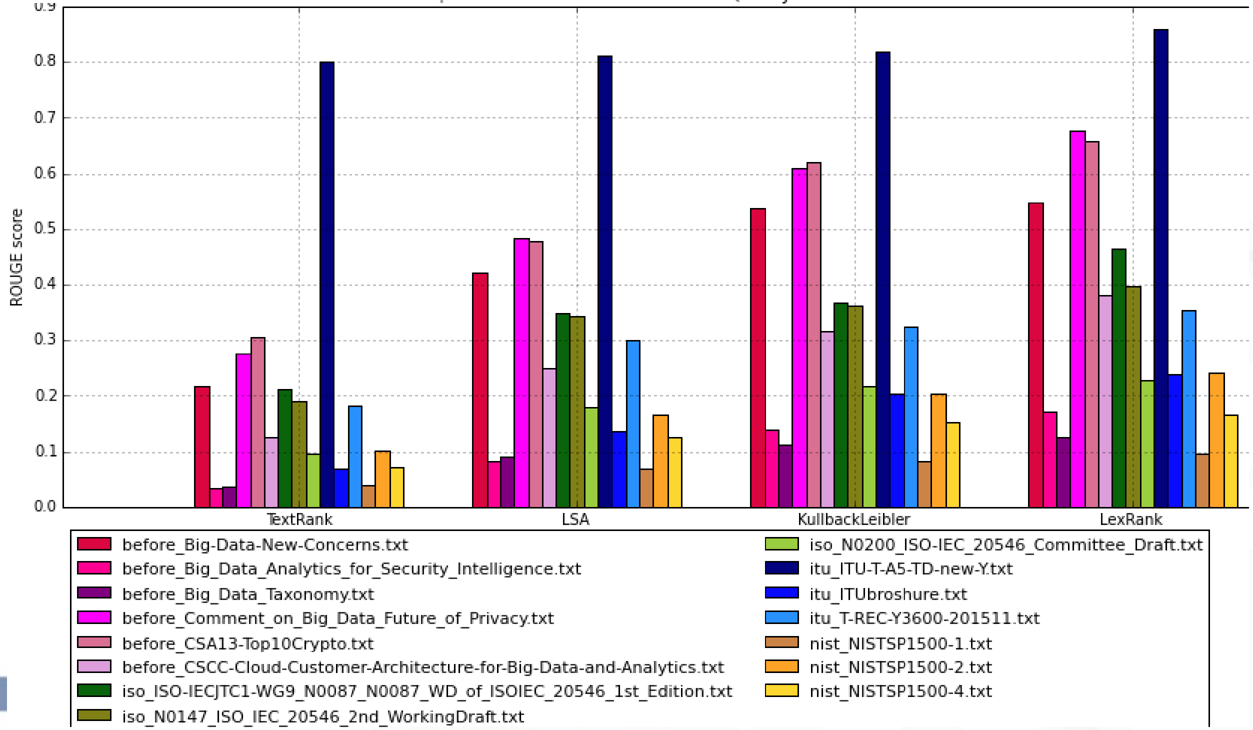
$$\text{ROUGE-N} = \frac{\displaystyle\sum_{S\in\{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\displaystyle\sum_{S\in\{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

$Count(gram_n)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries.

Denominator of the equation is the total sum of the number of n-grams occurring at the reference summary side.

Texts from schizophrenia area.

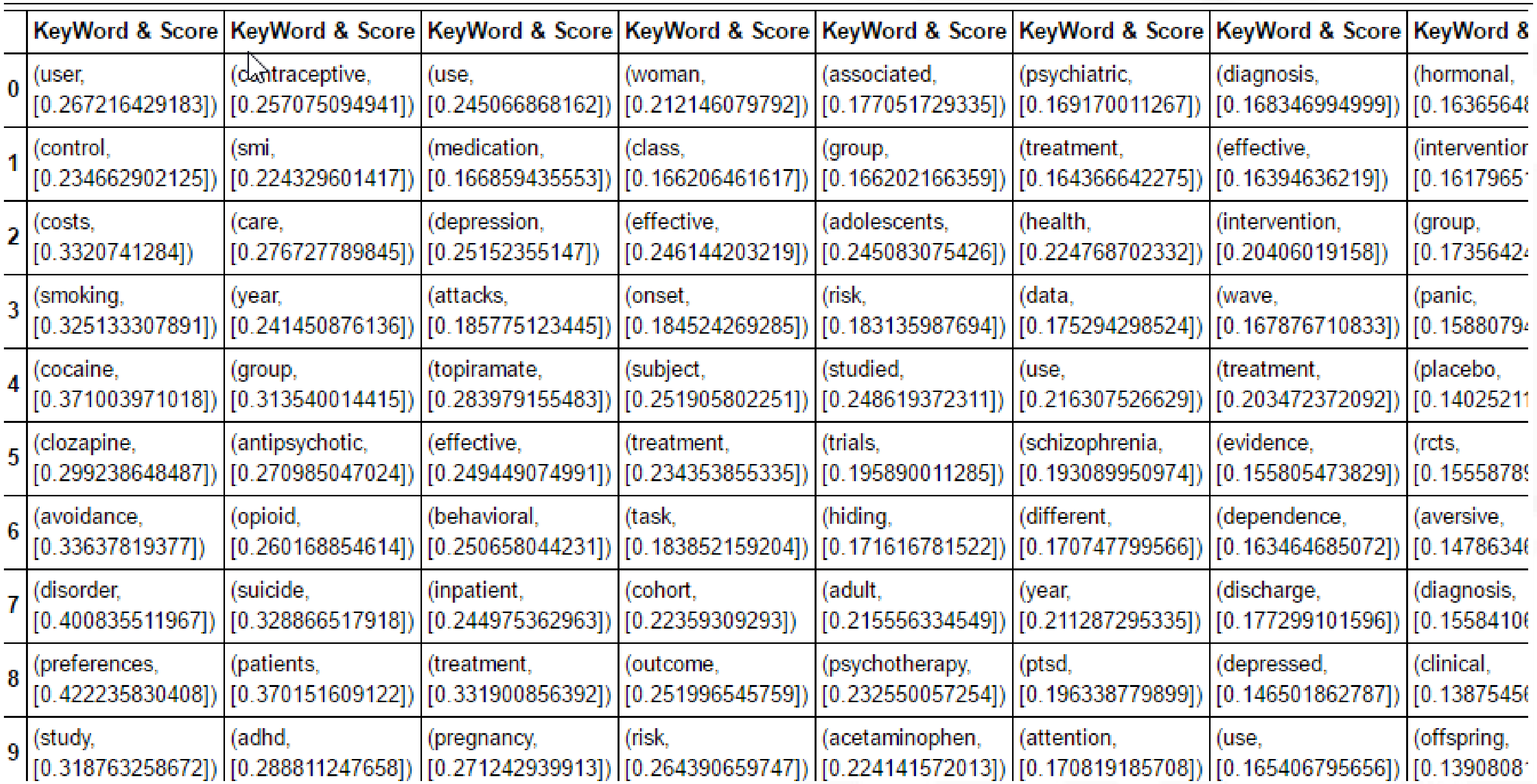




Texts from big data standardization.

Поиск ключевых слов, фраз в тексте (key words)

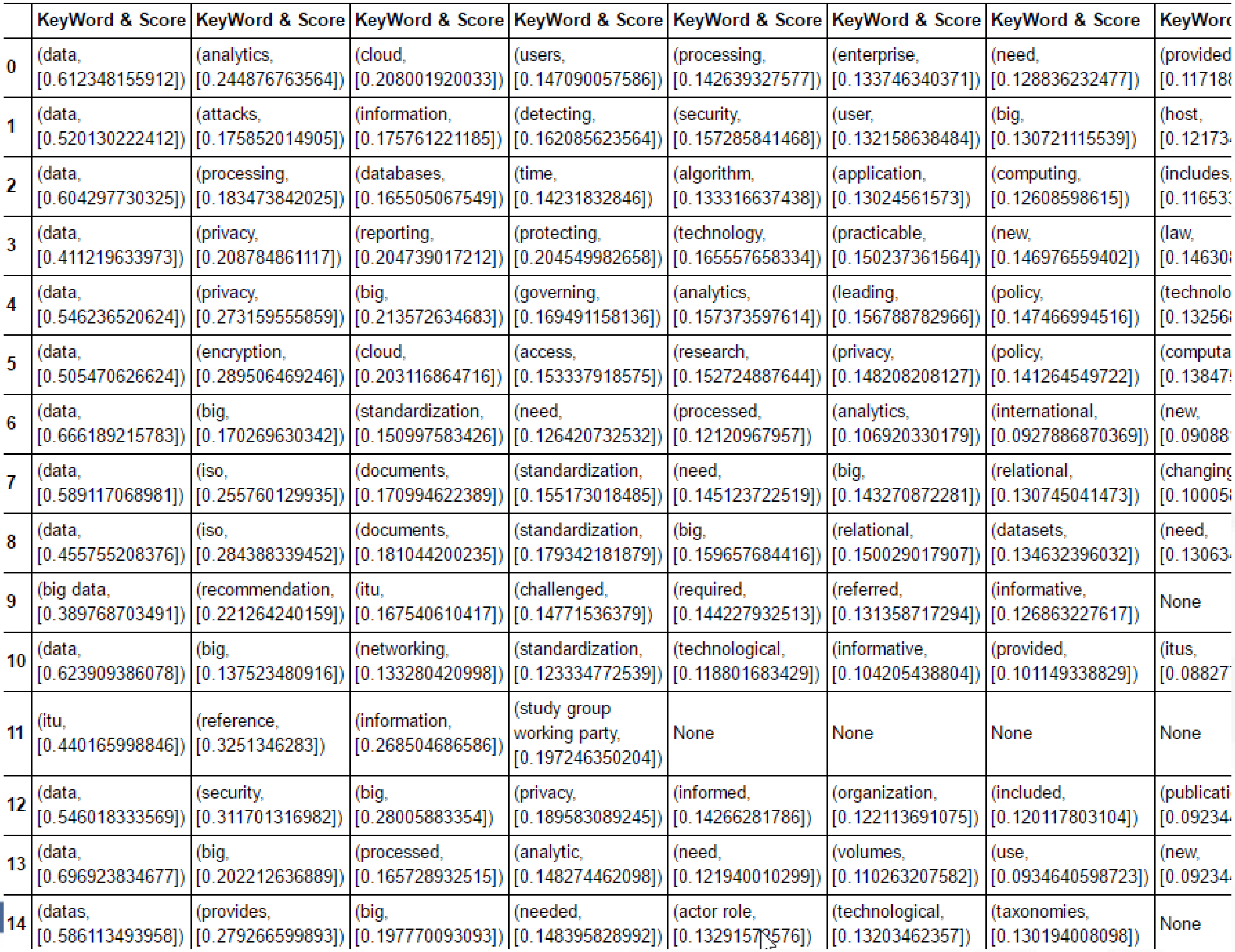
Dataset: texts from schizophrenia area.



	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord &
0	(user, [0.267216429183])	(contraceptive, [0.257075094941])	(use, [0.245066868162])	(woman, [0.212146079792])	(associated, [0.177051729335])	(psychiatric, [0.169170011267])	(diagnosis, [0.168346994999])	(hormonal, [0.1636564
1	(control, [0.234662902125])	(smi, [0.224329601417])	(medication, [0.166859435553])	(class, [0.166206461617])	(group, [0.166202166359])	(treatment, [0.164366642275])	(effective, [0.16394636219])	(interventio [0.1617965
2	(costs, [0.3320741284])	(care, [0.276727789845])	(depression, [0.25152355147])	(effective, [0.246144203219])	(adolescents, [0.245083075426])	(health, [0.224768702332])	(intervention, [0.20406019158])	(group, [0.173564
3	(smoking, [0.325133307891])	(year, [0.241450876136])	(attacks, [0.185775123445])	(onset, [0.184524269285])	(risk, [0.183135987694])	(data, [0.175294298524])	(wave, [0.167876710833])	(panic, [0.158807
4	(cocaine, [0.371003971018])	(group, [0.313540014415])	(topiramate, [0.283979155483])	(subject, [0.251905802251])	(studied, [0.248619372311])	(use, [0.216307526629])	(treatment, [0.203472372092])	(placebo, [0.140252
5	(clozapine, [0.299238648487])	(antipsychotic, [0.270985047024])	(effective, [0.249449074991])	(treatment, [0.234353855335])	(trials, [0.195890011285])	(schizophrenia, [0.193089950974])	(evidence, [0.155805473829])	(rcts, [0.155587
6	(avoidance, [0.33637819377])	(opioid, [0.260168854614])	(behavioral, [0.250658044231])	(task, [0.183852159204])	(hiding, [0.171616781522])	(different, [0.170747799566])	(dependence, [0.163464685072])	(aversive, [0.147863
7	(disorder, [0.400835511967])	(suicide, [0.328866517918])	(inpatient, [0.244975362963])	(cohort, [0.22359309293])	(adult, [0.215556334549])	(year, [0.211287295335])	(discharge, [0.177299101596])	(diagnosis, [0.155841
8	(preferences, [0.422235830408])	(patients, [0.370151609122])	(treatment, [0.331900856392])	(outcome, [0.251996545759])	(psychotherapy, [0.232550057254])	(ptsd, [0.196338779899])	(depressed, [0.146501862787])	(clinical, [0.138754
9	(study, [0.318763258672])	(adhd, [0.288811247658])	(pregnancy, [0.271242939913])	(risk, [0.264390659747])	(acetaminophen, [0.224141572013])	(attention, [0.170819185708])	(use, [0.165406795656])	(offspring, [0.139080

Поиск ключевых слов, фраз в тексте (key words)

Dataset: texts from big data standardization area.



	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWord & Score	KeyWorc
0	(data, [0.612348155912])	(analytics, [0.244876763564])	(cloud, [0.208001920033])	(users, [0.147090057586])	(processing, [0.142639327577])	(enterprise, [0.133746340371])	(need, [0.128836232477])	(provided [0.11718
1	(data, [0.520130222412])	(attacks, [0.175852014905])	(information, [0.175761221185])	(detecting, [0.162085623564])	(security, [0.157285841468])	(user, [0.132158638484])	(big, [0.130721115539])	(host, [0.12173
2	(data, [0.604297730325])	(processing, [0.183473842025])	(databases, [0.165505067549])	(time, [0.14231832846])	(algorithm, [0.133316637438])	(application, [0.13024561573])	(computing, [0.12608598615])	(includes, [0.11653
3	(data, [0.411219633973])	(privacy, [0.208784861117])	(reporting, [0.204739017212])	(protecting, [0.204549982658])	(technology, [0.165557658334])	(practicable, [0.150237361564])	(new, [0.146976559402])	(law, [0.14630
4	(data, [0.546236520624])	(privacy, [0.273159555859])	(big, [0.213572634683])	(governing, [0.169491158136])	(analytics, [0.157373597614])	(leading, [0.156788782966])	(policy, [0.147466994516])	(technolo [0.13256
5	(data, [0.505470626624])	(encryption, [0.289506469246])	(cloud, [0.203116864716])	(access, [0.153337918575])	(research, [0.152724887644])	(privacy, [0.148208208127])	(policy, [0.141264549722])	(computa [0.13847
6	(data, [0.666189215783])	(big, [0.170269630342])	(standardization, [0.150997583426])	(need, [0.126420732532])	(processed, [0.12120967957])	(analytics, [0.106920330179])	(international, [0.0927886870369])	(new, [0.09088
7	(data, [0.589117068981])	(iso, [0.255760129935])	(documents, [0.170994622389])	(standardization, [0.155173018485])	(need, [0.145123722519])	(big, [0.143270872281])	(relational, [0.130745041473])	(changing [0.10005
8	(data, [0.455755208376])	(iso, [0.284388339452])	(documents, [0.181044200235])	(standardization, [0.179342181879])	(big, [0.159657684416])	(relational, [0.150029017907])	(datasets, [0.134632396032])	(need, [0.13063
9	(big data, [0.389768703491])	(recommendation, [0.221264240159])	(itu, [0.167540610417])	(challenged, [0.14771536379])	(required, [0.144227932513])	(referred, [0.131358717294])	(informative, [0.126863227617])	None
10	(data, [0.623909386078])	(big, [0.137523480916])	(networking, [0.133280420998])	(standardization, [0.123334772539])	(technological, [0.118801683429])	(informative, [0.104205438804])	(provided, [0.101149338829])	(itus, [0.08827
11	(itu, [0.440165998846])	(reference, [0.3251346283])	(information, [0.268504686586])	(study group working party, [0.197246350204])	None	None	None	None
12	(data, [0.546018333569])	(security, [0.311701316982])	(big, [0.28005883354])	(privacy, [0.189583089245])	(informed, [0.14266281786])	(organization, [0.122113691075])	(included, [0.120117803104])	(publicati [0.09234
13	(data, [0.696923834677])	(big, [0.202212636889])	(processed, [0.165728932515])	(analytic, [0.148274462098])	(need, [0.121940010299])	(volumes, [0.110263207582])	(use, [0.0934640598723])	(new, [0.09234
14	(datas, [0.586113493958])	(provides, [0.279266599893])	(big, [0.197770093093])	(needed, [0.148395828992])	(actor role, [0.1329157[illegible]576])	(technological, [0.13203462357])	(taxonomies, [0.130194008098])	None