

Search Engine

Paweł Konopka

1. Dataset

Dataset consists entirely of English **Wikipedia articles** somehow connected to **history**. Their contents and titles were downloaded using wikipedia python library as queries with following keywords:

```
KEYWORDS = [  
    "world history",  
    "human history",  
    "evolution",  
    "homo",  
    "medieval",  
    "middle ages",  
    "renaissance",  
    "war",  
    "battle",  
    "revolution",  
    "invasion",  
    "europe",  
    "asia",  
    "america",  
    "africa",  
    "poland",  
    "european union",  
    "united states",  
    "united kingdom",  
    "history",  
    "archeology",  
    "leader",  
    "king",  
    "queen",  
    "emperor",  
    "general",  
]
```

In total **10,178 articles** were collected. Even though there is no obstacle to download any number of wikipedia pages, I was forced to stop here because of RAM limitations on my PC.

The whole code regarding Wikipedia crawling is located in **wikidata/wiki_download.py**.

2. Data pre-processing

a. Splitting

In order to efficiently use the dataset, it was required to pre-process its content. First step was to split each article into words (called later terms). I decided to simply split each article string with any of the following characters as separator:

```
re.split(r'[\t\n\r\v\f,.!:\;\-\=\/\(\)\[\]\\'\"'], line)
```

The fact that helped a lot is that Wikipedia articles are free of HTML tags, that could be a problem otherwise. After splitting only those words are taken into consideration, whose length is greater than one and all characters are alphabetic. This way all numbers are ignored, which may not be the best solution (for example for differentiating between World War I and World War II). Next all words are being transformed to lowercase.

b. Stemming

Word stemming was achieved by using Porter Stemmer from **nltk** Python library.

c. Stop words removal

Words from **nltk** English stop words list were manually removed from each article processing.

Functions implementing aforementioned functionalities can be found in `engine/text_analysys.py`.

3. Creating vocabulary

Vocabulary was created by counting all usages of each unique term (after pre-processing) in all articles. Vocabulary size was arbitrarily set to **5,000**. Again RAM limitations were the only obstacle in increasing this number. All other words than the 5,000 most used in articles after pre-processing were simply ignored later on by the search engine.

4. Term frequency

Term frequency (**tf**) is given as a share each term holds in each article. Formally we have:

$$tf(t, d) = \frac{count(t, d)}{length(d)}$$

Where: t - term, d - document

Term frequency value is always in range: **[0, 1]**.

In Python it is represented as a list of dictionaries. List index corresponds to article index of being read. Dictionary keys are terms, and values their frequencies for given documents.

5. Document frequency

Document frequency (**df**) is given as a number of articles in which certain terms is used (at least once). The equation goes as follows:

$$df(t) = \sum_{d \in D} \frac{\mathbb{1}(\text{count}(t, d) \geq 1)}{|D|}$$

Where: D – set of all documents

Term frequency value is also always in range: $[0, 1]$.

In python it is represented as a dictionary where each term is a key, and its document frequency a value.

6. Inverse document frequency

Inverse document frequency (**idf**) is given as:

$$idf(t) = \log \frac{|D|}{df(t) + 1}$$

Since df is a measure of how often does certain word appear in the text, higher idf value means that given term is quite rare and unique. Logarithmic function is used to smoothen this out, which tends to present better results.

Dictionary Python representation remains the same.

7. TF-IDF: Inverse Document Frequency

The final value that will be taken into consideration during computation is **tf-idf**. It is given by a formula:

$$tf-idf(t, d) = tf(t, d) \cdot idf(t)$$

It is basically tf value (how often does certain term appear in certain document) multiplied by its “uniqueness” (idf value).

In the code $tf-idf$ is initially represented as a list of dictionaries (same as tf). Later on it is being transformed into $|T| \times |D|$ sparse matrix, where T stands for set of all terms (vocabulary). Each matrix column was also normalized

8. Singular Value Decomposition

Using Singular Value Decomposition we are able to transfer matrix M as product of three matrices:

$$M = U\Sigma V^T$$

Truncation of SVD allows us to reduce matrix size. I used ready implementation from **sklearn** Python library. It returns two matrices: $(U\Sigma)_{|T|\times k}$ and $V^T_{k\times |D|}$.

Since it is an approximation, their product will not be exactly equal to *tf-idf* matrix.

9. Frontend layout

Frontend layout was added using **Django Python framework** with plain **HTML** and **CSS**.

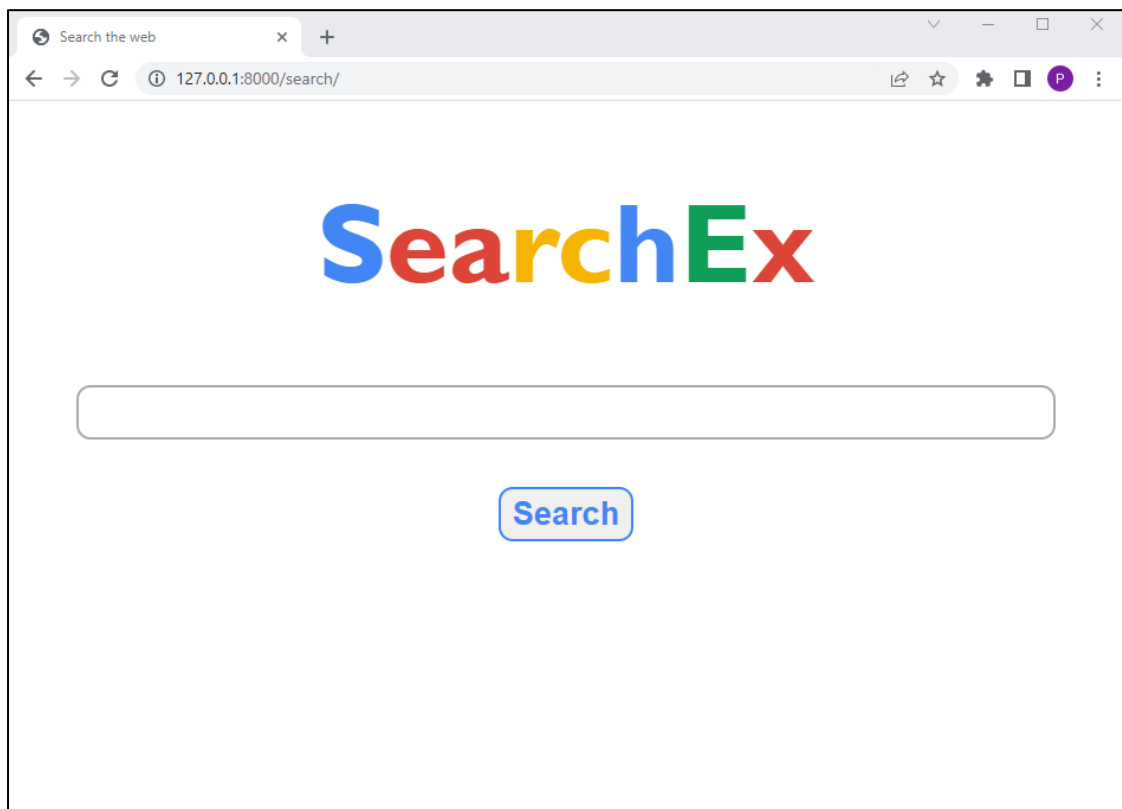


Figure 1: Search page graphical view

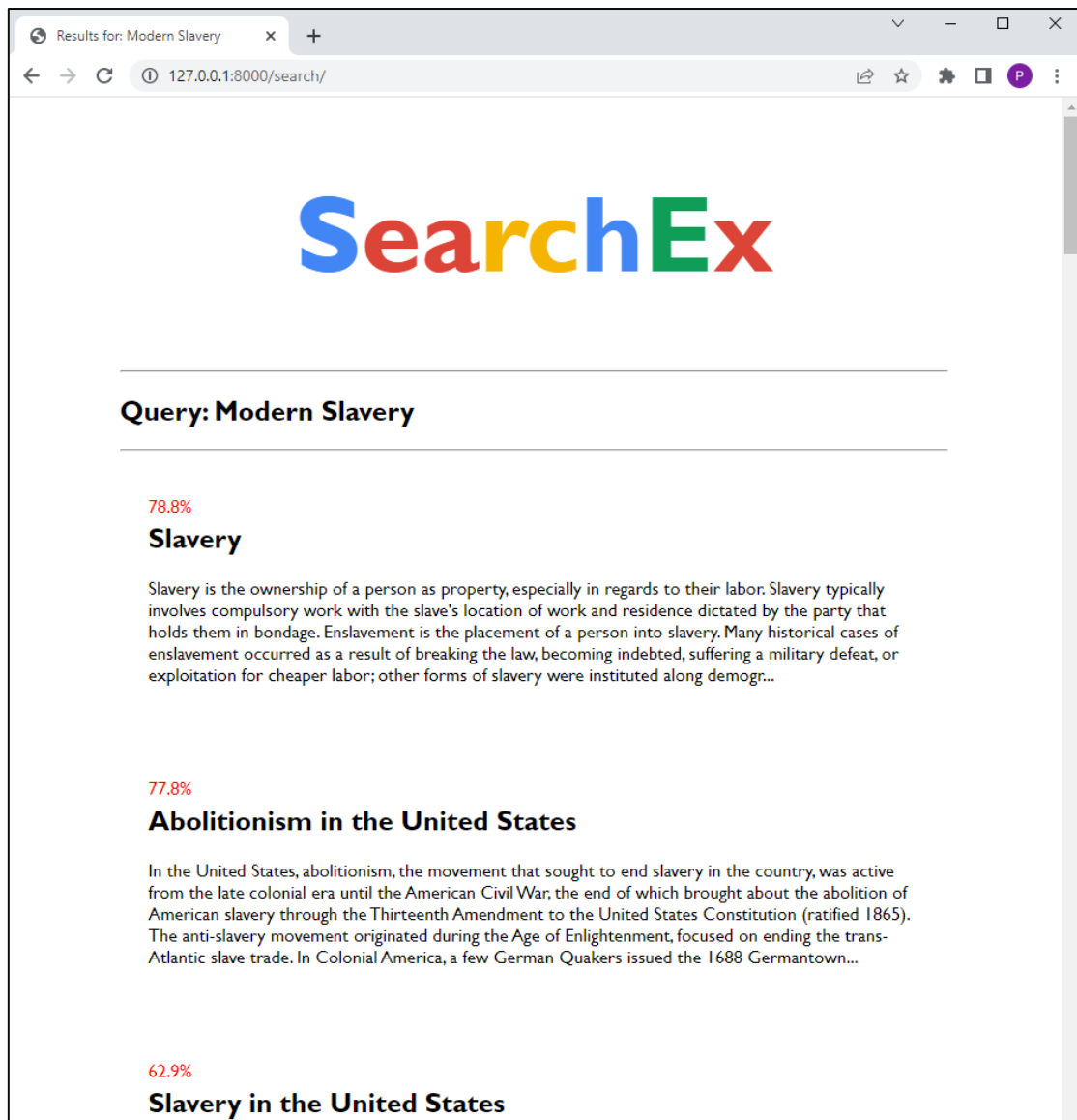


Figure 2: Search page graphical view

10.Result comparison using different parameters

a. Query: Ukrainian people and culture

i. No SVD search

Query: Ukrainian people and culture	
78.2%	Ukrainian People's Republic The Ukrainian People's Republic (UPR) was a short-lived state in Eastern Europe. Prior to its proclamation, the Central Council of Ukraine was elected in March 1917 as a result of the February Revo...
77.5%	Ukrainians in Poland Ukrainians in Poland have various legal statuses: ethnic minority, temporary and permanent residents, and refugees. According to the Polish census of 2011, the Ukrainian minority in Poland was comp...
77.2%	Ukrainians in the United Kingdom Ukrainians in the United Kingdom consist mainly of British citizens of Ukrainian descent. == History == In Manchester, the first documented evidence of Ukrainians was an entry in the Allens Regs...
73.6%	Ukrainian War of Independence The Ukrainian War of Independence was a series of conflicts involving many adversaries that lasted from 1917 to 1921 and resulted in the establishment and development of a Ukrainian republic, most ...
72.7%	Ukrainian Americans Ukrainian Americans (Ukrainian: українські американці, romanized: Ukraïnski amerykantsi) are Americans who are of Ukrainian ancestry. According to U.S. census estimates, in 2021 there were 1,017,5...
69.9%	Casualties of the Russo-Ukrainian War Casualties in the Russo-Ukrainian War included six deaths during the 2014 annexation of Crimea by the Russian Federation, 14,200–14,400 military and civilian deaths during the war in Donbas (2014–2...
65.0%	Executed Renaissance The Executed Renaissance (or "Red Renaissance", Ukrainian: Розстріляне відродження, Червоний ренесанс, romanized: Rozstriliane vidrodzhennia, Chervonyi renesans) is a term used to describe the gene...
64.5%	Andriy Melnyk (officer) Andriy Atanasovich Melnyk (Ukrainian: Андрій Атанасович Мельник; 12 December 1890 – 1 November 1964) was a Ukrainian military and political leader. == Life == Melnyk was born near Drohobych, Hal...
61.3%	2022–present Ukrainian refugee crisis An ongoing refugee crisis began in Europe in late February 2022 after Russia's invasion of Ukraine. Over 8.2 million refugees fleeing Ukraine have been recorded across Europe, while an estimated 8 ...
58.8%	Donetsk People's Republic The Donetsk People's Republic (Russian: Донецкая Народная Республика, tr. Donetskaya Narodnaya Respublika, IPA: [dɐˈnʲetskɔjə nɐˈrodnɔjə rʲɪˈspublʲɪkə]; abbreviated as DPR or DNR, Russian: ДНР) is ...

Figure 3: Results for query “Ukrainian people and culture” with SVD turned off

ii. SVD search, k = 100

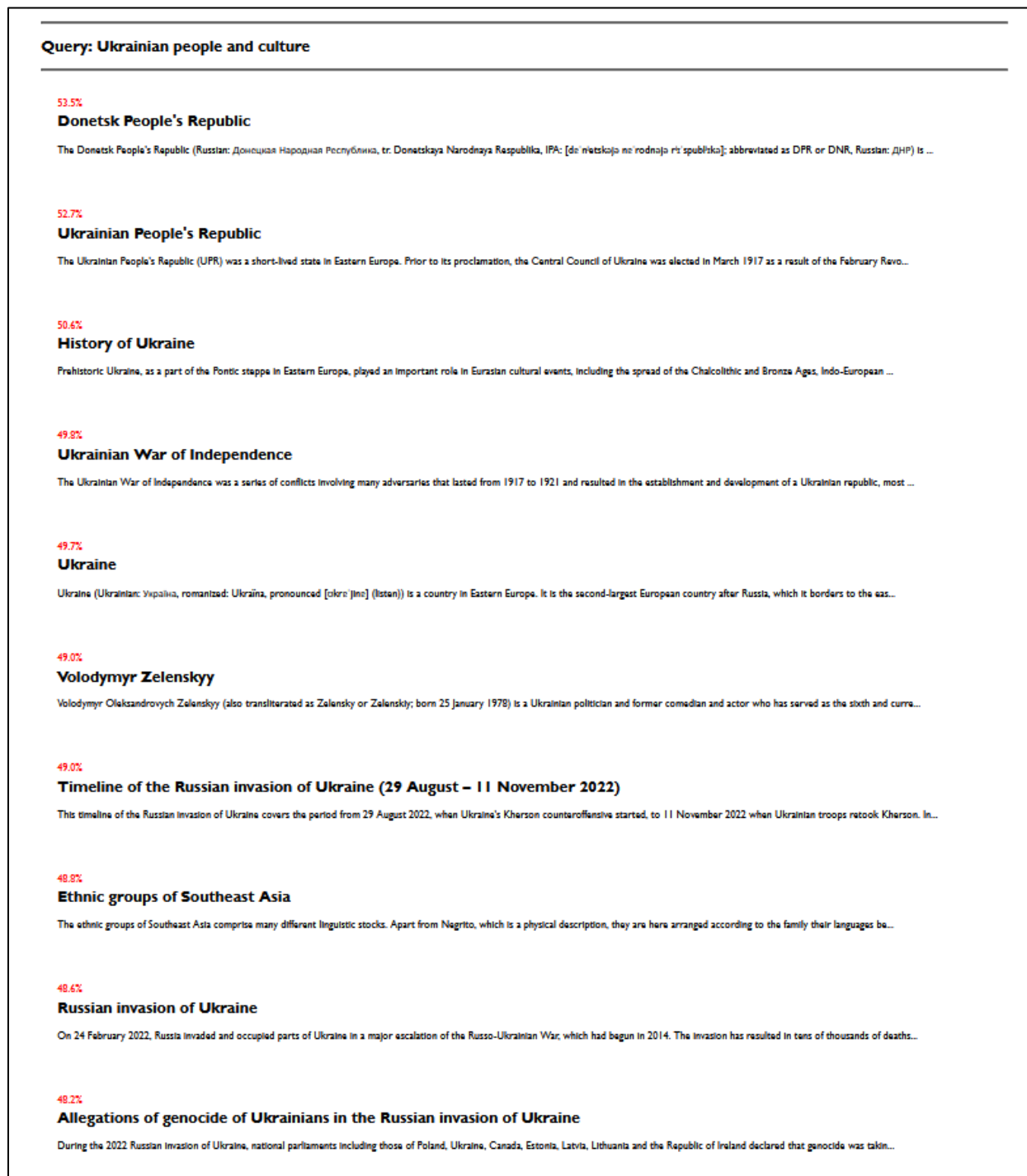


Figure 4: Results for query “Ukrainian people and culture” with SVD k=100

iii. SVD search, k = 1000

Query: Ukrainian people and culture	
78.2%	Ukrainian People's Republic The Ukrainian People's Republic (UPR) was a short-lived state in Eastern Europe. Prior to its proclamation, the Central Council of Ukraine was elected in March 1917 as a result of the February Rev...
77.4%	Ukrainians in Poland Ukrainians in Poland have various legal statuses: ethnic minority, temporary and permanent residents, and refugees. According to the Polish census of 2011, the Ukrainian minority in Poland was comp...
77.1%	Ukrainians in the United Kingdom Ukrainians in the United Kingdom consist mainly of British citizens of Ukrainian descent. == History == In Manchester, the first documented evidence of Ukrainians was an entry in the Aliens Regis...
73.5%	Ukrainian War of Independence The Ukrainian War of Independence was a series of conflicts involving many adversaries that lasted from 1917 to 1921 and resulted in the establishment and development of a Ukrainian republic, most ...
72.6%	Ukrainian Americans Ukrainian Americans (Ukrainian: Українські американці, romanized: Ukraïnskî amerykantsi) are Americans who are of Ukrainian ancestry. According to U.S. census estimates, in 2021 there were 1,017,5...
69.8%	Casualties of the Russo-Ukrainian War Casualties in the Russo-Ukrainian War included six deaths during the 2014 annexation of Crimea by the Russian Federation, 14,200–14,400 military and civilian deaths during the war in Donbas (2014–2...
65.0%	Executed Renaissance The Executed Renaissance (or "Red Renaissance", Ukrainian: Розстріляне відродження, Червоний ренесанс, romanized: Rozstriliane vidrodzhennia, Chervomyi renesans) is a term used to describe the gene...
64.5%	Andriy Melnyk (officer) Andriy Atanasovych Melnyk (Ukrainian: Андрій Атанасович Мельник; 12 December 1890 – 1 November 1964) was a Ukrainian military and political leader. == Life == Melnyk was born near Drohobych, Hal...
61.2%	2022–present Ukrainian refugee crisis An ongoing refugee crisis began in Europe in late February 2022 after Russia's invasion of Ukraine. Over 8.2 million refugees fleeing Ukraine have been recorded across Europe, while an estimated 8 ...
58.7%	Donetsk People's Republic The Donetsk People's Republic (Russian: Донецкая Народная Республика, tr. Donetskaya Narodnaya Respublika, IPA: [dɐˈnʲetskɔjə nɐˈrodnɔjə rʲɪˈspublʲɪkə]; abbreviated as DPR or DNR, Russian: ДНР) is ...

Figure 5: Results for query “Ukrainian people and culture” with SVD k=1000

b. Query: German holy military order

i. No SVD search

Query: German holy military order	
68.9%	Teutonic Order The Teutonic Order is a Catholic religious institution founded as a military society c. 1190 in Acre, Kingdom of Jerusalem. The Order of Brothers of the German House of Saint Mary in Jerusalem was ...
64.8%	Order of the October Revolution The Order of the October Revolution (Russian: Орден Октябрьской Революции, Orden Oktyabr'skoy Revolyutsii) was instituted on October 31, 1967, in time for the 50th anniversary of the October Revolu...
63.0%	Master of the Order of Preachers The Master of the Order of Preachers is the Superior General of the Order of Preachers, commonly known as the Dominicans. The Master of the Order of Preachers is ex officio Grand Chancellor of the P...
61.5%	Order of battle In modern use, the order of battle of an armed force participating in a military operation or campaign shows the hierarchical organization, command structure, strength, disposition of personnel, an...
61.0%	Orders, decorations, and medals of Poland The following is a list of medals, awards and decorations in use in Poland. Most of them are awarded by the Polish Army, but some of them are civilian decorations that may be worn by the military p...
56.8%	Orders, decorations, and medals of the United Kingdom In the United Kingdom and the British Overseas Territories, personal bravery, achievement, or service are rewarded with honours. The honours system consists of three types of award: Honours are us...
55.4%	Supreme Order of the Renaissance The Supreme Order of the Renaissance (Arabic: وسام النهضة, romanized: wistim an-nahda, "Medal of the Nahda") is the second knighthood order of the Kingdom of Jordan. == History == The order was L...
54.8%	Queen Silvia of Sweden Silvia (born Silvia Renate Sommerlath; 23 December 1943) is Queen of Sweden as the wife of King Carl XVI Gustaf. She has held this title since her marriage to Carl Gustaf in 1976. The king and quee...
52.0%	Orders of precedence in the United Kingdom The order of precedence in the United Kingdom is the sequential hierarchy for Peers of the Realm, officers of state, senior members of the clergy, holders of the various Orders of Chivalry and othe...
49.7%	Queen Paola of Belgium Paola (née Donna Paola Ruffo di Calabria; born 11 September 1937) is a member of the Belgian royal family who was Queen of the Belgians during the reign of her husband, King Albert II, from 9 Augu...

Figure 6: Results for query “German holy military order” with SVD turned off

ii. SVD search, k = 100

Query: German holy military order	
54.2%	Teutonic Order The Teutonic Order is a Catholic religious institution founded as a military society c. 1190 in Acre, Kingdom of Jerusalem. The Order of Brothers of the German House of Saint Mary in Jerusalem was ...
49.2%	Orders, decorations, and medals of Poland The following is a list of medals, awards and decorations in use in Poland. Most of them are awarded by the Polish Army, but some of them are civilian decorations that may be worn by the military p...
49.2%	Order of the October Revolution The Order of the October Revolution (Russian: Орден Октябрьской Революции, Orden Oktyabr'skoy Revolyutsii) was instituted on October 31, 1967, in time for the 50th anniversary of the October Revolu...
49.0%	Queen Silvia of Sweden Silvia (born Silvia Renata Sommerlath; 23 December 1943) is Queen of Sweden as the wife of King Carl XVI Gustaf. She has held this title since her marriage to Carl Gustaf in 1976. The king and quee...
47.8%	Orders, decorations, and medals of the United Kingdom In the United Kingdom and the British Overseas Territories, personal bravery, achievement, or service are rewarded with honours. The honours system consists of three types of award: Honours are us...
47.5%	Master of the Order of Preachers The Master of the Order of Preachers is the Superior General of the Order of Preachers, commonly known as the Dominicans. The Master of the Order of Preachers is ex officio Grand Chancellor of the P...
47.3%	Supreme Order of the Renaissance The Supreme Order of the Renaissance (Arabic: وسام النهضة, romanized: wisaam an-nahda, "Medal of the Nahda") is the second knighthood order of the Kingdom of Jordan. == History == The order was l...
44.6%	Order of battle In modern use, the order of battle of an armed force participating in a military operation or campaign shows the hierarchical organization, command structure, strength, disposition of personnel, an...
43.0%	Queen Paola of Belgium Paola (née Donna Paola Ruffo di Calabria; born 11 September 1937) is a member of the Belgian royal family who was Queen of the Belgians during the reign of her husband, King Albert II, from 9 Augus...

Figure 7: Results for query “German holy military order” with SVD k=1000

iii. SVD search, k = 1000

Query: German holy military order	
68.4%	Teutonic Order The Teutonic Order is a Catholic religious institution founded as a military society c. 1190 in Acre, Kingdom of Jerusalem. The Order of Brothers of the German House of Saint Mary in Jerusalem was ...
64.5%	Order of the October Revolution The Order of the October Revolution (Russian: Орден Октябрьской Революции, Orden Oktyabr'skoy Revolyutsii) was instituted on October 31, 1967, in time for the 50th anniversary of the October Revolu...
62.9%	Master of the Order of Preachers The Master of the Order of Preachers is the Superior General of the Order of Preachers, commonly known as the Dominicans. The Master of the Order of Preachers is ex officio Grand Chancellor of the P...
61.3%	Order of battle In modern use, the order of battle of an armed force participating in a military operation or campaign shows the hierarchical organization, command structure, strength, disposition of personnel, an...
60.5%	Orders, decorations, and medals of Poland The following is a list of medals, awards and decorations in use in Poland. Most of them are awarded by the Polish Army, but some of them are civilian decorations that may be worn by the military p...
56.9%	Orders, decorations, and medals of the United Kingdom In the United Kingdom and the British Overseas Territories, personal bravery, achievement, or service are rewarded with honours. The honours system consists of three types of award: Honours are us...
55.8%	Supreme Order of the Renaissance The Supreme Order of the Renaissance (Arabic: وسام النهضة, romanized: wṣām an-nahḍa, "Medal of the Nahḍa") is the second knighthood order of the Kingdom of Jordan. == History == The order was L...
55.4%	Queen Silvia of Sweden Silvia (born Silvia Renata Sommerlath; 23 December 1943) is Queen of Sweden as the wife of King Carl XVI Gustaf. She has held this title since her marriage to Carl Gustaf in 1976. The king and quee...
52.5%	Orders of precedence in the United Kingdom The order of precedence in the United Kingdom is the sequential hierarchy for Peers of the Realm, officers of state, senior members of the clergy, holders of the various Orders of Chivalry and othe...

Figure 8: Results for query “German holy military order” with SVD k=1000

11. Subjective comparison between different search parameters

Using shown above examples and more testing SVD search with high k values seem to be giving very similar results to no-SVD search. In particular circumstances SVD may prove to better “understand” query meaning. On the other hand, low k values searches provide rather unrelated results.