FAKULTA
INFORMATIKY
Masarykova univerzita

Vektorové reprezentace ve vyhledávání znalostí

Vector Space Representations in Information Retrieval

Vít Novotný
witiko@mail.muni.cz

6. února 2017

Obsah

FAKULTA
INFORMATIKY
Masarykova univerzita

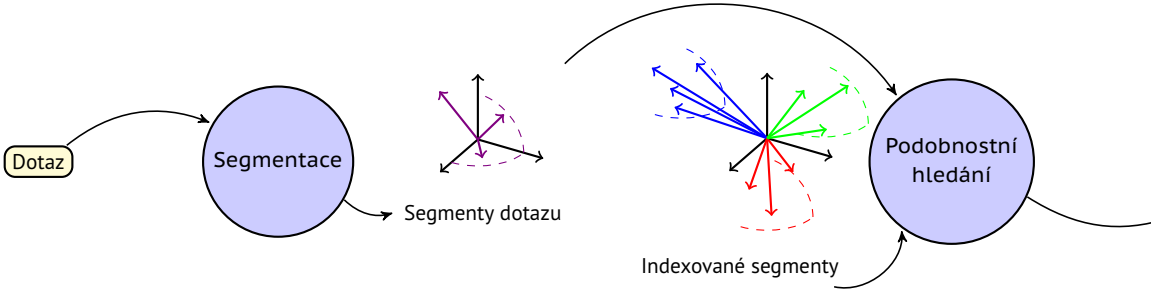
Úvod

- V rámci výzkumné skupiny Math Information Retrieval (MIR) jsem se ve spolupráci s firmou RaRe Technologies zúčastnil třetího kola programu **TA ČR Omega**.

FAKULTA INFORMATIKY
Masarykova univerzita

Úvod

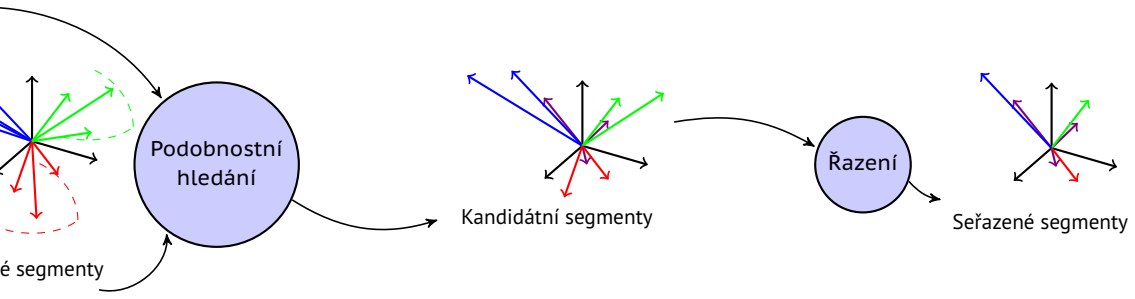
- V rámci výzkumné skupiny Math Information Retrieval (MIR) jsem se ve spolupráci s firmou RaRe Technologies zúčastnil třetího kola programu TA ČR Omega.
- Cíl byl vyvinout **segmentující vyhledávač** nestrukturovaných textových dokumentů:



FAKULTA INFORMATIKY
Masarykova univerzita

Úvod

- V rámci výzkumné skupiny Math Information Retrieval (MIR) jsem se ve spolupráci s firmou RaRe Technologies zúčastnil třetího kola programu TA ČR Omega.
- Cíl byl vyvinout **segmentující vyhledávač** nestrukturovaných textových dokumentů:



Úvod

- V rámci výzkumné skupiny Math Information Retrieval (MIR) jsem se ve spolupráci s firmou RaRe Technologies zúčastnil třetího kola programu TA ČR Omega.
- Cíl byl vyvinout segmentující vyhledávač nestrukturovaných textových dokumentů.
- V rámci projektu jsem dostal možnost prezentovat náš výzkum na ACL 2017.

FAKULTA INFORMATIKY
Masarykova univerzita

Úvod

- V rámci výzkumné skupiny Math Information Retrieval (MIR) jsem se ve spolupráci s firmou RaRe Technologies zúčastnil třetího kola programu TA ČR Omega.
- Cíl byl vyvinout segmentující vyhledávač nestrukturovaných textových dokumentů.
- V rámci projektu jsem dostal možnost prezentovat náš výzkum na ACL 2017.
- V návaznosti na výzkum prezentovaný na ACL 2017 jsem provedl **sérii experimentů**, které jsou předmětem této diplomové práce:

Úvod

- V rámci výzkumné skupiny Math Information Retrieval (MIR) jsem se ve spolupráci s firmou RaRe Technologies zúčastnil třetího kola programu TA ČR Omega.
- Cíl byl vyvinout segmentující vyhledávač nestrukturovaných textových dokumentů.
- V rámci projektu jsem dostal možnost prezentovat náš výzkum na ACL 2017.
- V návaznosti na výzkum prezentovaný na ACL 2017 jsem provedl **sérii experimentů**, které jsou předmětem této diplomové práce:
  1. U vyhledávačů, které musí vždy navrátit celé dokumenty a nikoliv pouze segmenty, lze **agregací nalezených segmentů** zlepšit kvalitu výsledků oproti hledání bez segmentace.

FAKULTA INFORMATIKY
Masarykova univerzita

Úvod

- V rámci výzkumné skupiny Math Information Retrieval (MIR) jsem se ve spolupráci s firmou RaRe Technologies zúčastnil třetího kola programu TA ČR Omega.
- Cíl byl vyvinout segmentující vyhledávač nestrukturovaných textových dokumentů.
- V rámci projektu jsem dostal možnost prezentovat náš výzkum na ACL 2017.
- V návaznosti na výzkum prezentovaný na ACL 2017 jsem provedl **sérii experimentů**, které jsou předmětem této diplomové práce:
  1. U vyhledávačů, které musí vždy navrátit celé dokumenty a nikoliv pouze segmenty, lze agregací nalezených segmentů zlepšit kvalitu výsledků oproti hledání bez segmentace.
  2. Rozšířením standardního vektorového modelu o neortogonalitu mezi bázovými vektory lze **modelovat synonymitu slov** a docílit dalšího zlepšení kvality výsledků.

FAKULTA
INFORMATIKY
Masarykova univerzita

Datová sada

- V rámci obou experimentů jsem využil datovou sadu pro úlohu 3 (komunitní poradny) z ročníků 2016 a 2017 soutěže SemEval.

Datová sada

- V rámci obou experimentů jsem využil datovou sadu pro úlohu 3 (komunitní poradny) z ročníků 2016 a 2017 soutěže SemEval.
- Datové sady pro podúlohu 3a obsahují **vlákna s otázkou a prvními deseti komentáři** spolu s anotací, jestli **je komentář relevantní vůči otázce**.

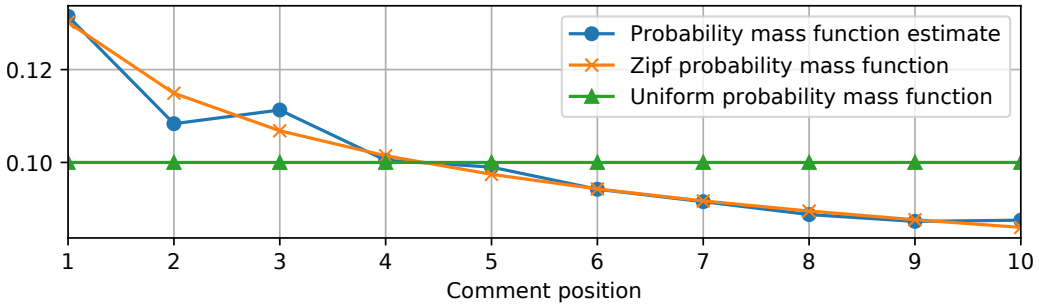
Datová sada

- V rámci obou experimentů jsem využil datovou sadu pro úlohu 3 (komunitní poradny) z ročníků 2016 a 2017 soutěže SemEval.
- Datové sady pro podúlohu 3a obsahují **vlákna s otázkou a prvními deseti komentáři** spolu s anotací, jestli **je komentář relevantní vůči otázce**.
  - Mike Godwin roku 1991 formuloval empirické pravidlo, že „s rostoucí délkou UseNetové diskuze se pravděpodobnost přirovnání zmiňujícího nacisty nebo Hitlera blíží k jedné.“

FAKULTA INFORMATIKY
Masarykova univerzita

Datová sada

- V rámci obou experimentů jsem využil datovou sadu pro úlohu 3 (komunitní poradny) z ročníků 2016 a 2017 soutěže SemEval.
- Datové sady pro podúlohu 3a obsahují **vlákna s otázkou a prvními deseti komentáři** spolu s anotací, jestli **je komentář relevantní vůči otázce**.
  - Mike Godwin roku 1991 formuloval empirické pravidlo, že „s rostoucí délkou UseNetové diskuze se pravděpodobnost přirovnání zmiňujícího nacisty nebo Hitlera blíží k jedné.“
  - Na základě tohoto pravidla jsem formuloval a vyvrátil hypotézu, že **pravděpodobnost výskytu relevantních komentářů na jednotlivých pozicích je rovnoměrná**.



Obrázek: Odhad pravděpodobnostní funkce $P(\text{na pozici } i \mid \text{relevantní})$ vyobrazený modře spolu s pravděpodobnostními funkcemi Zipfova (oranžový graf) a rovnoměrného rozdělení (zelený graf).

Datová sada

- V rámci obou experimentů jsem využil datovou sadu pro úlohu 3 (komunitní poradny) z ročníků 2016 a 2017 soutěže SemEval.
- Datové sady pro podúlohu 3a obsahují **vlákna s otázkou a prvními deseti komentáři** spolu s anotací, jestli **je komentář relevantní vůči otázce**.
  - Mike Godwin roku 1991 formuloval empirické pravidlo, že „s rostoucí délkou UseNetové diskuze se pravděpodobnost přirovnání zmiňujícího nacisty nebo Hitlera blíží k jedné.“
  - Na základě tohoto pravidla jsem formuloval a vyvrátil hypotézu, že **pravděpodobnost výskytu relevantních komentářů na jednotlivých pozicích je rovnoměrná**.

Datová sada

- V rámci obou experimentů jsem využil datovou sadu pro úlohu 3 (komunitní poradny) z ročníků 2016 a 2017 soutěže SemEval.
- Datové sady pro podúlohu 3a obsahují **vlákna s otázkou a prvními deseti komentáři** spolu s anotací, jestli **je komentář relevantní vůči otázce**.
  - Mike Godwin roku 1991 formuloval empirické pravidlo, že „s rostoucí délkou UseNetové diskuze se pravděpodobnost přirovnání zmiňujícího nacisty nebo Hitlera blíží k jedné.“
  - Na základě tohoto pravidla jsem formuloval a vyvrátil hypotézu, že pravděpodobnost výskytu relevantních komentářů na jednotlivých pozicích je rovnoměrná.
- Datové sady pro podúlohu 3b obsahují dotazy a **pro každý dotaz deset vláken** spolu s anotací, jestli se **vlákno týká dotazu**. Vlákna řadíme podle podobnosti k dotazu.

Datová sada

- V rámci obou experimentů jsem využil datovou sadu pro úlohu 3 (komunitní poradny) z ročníků 2016 a 2017 soutěže SemEval.
- Datové sady pro podúlohu 3a obsahují **vlákna s otázkou a prvními deseti komentáři** spolu s anotací, jestli **je komentář relevantní vůči otázce**.
  - Mike Godwin roku 1991 formuloval empirické pravidlo, že „s rostoucí délkou UseNetové diskuze se pravděpodobnost přirovnání zmiňujícího nacisty nebo Hitlera blíží k jedné."
  - Na základě tohoto pravidla jsem formuloval a vyvrátil hypotézu, že pravděpodobnost výskytu relevantních komentářů na jednotlivých pozicích je rovnoměrná.
- Datové sady pro podúlohu 3b obsahují dotazy a **pro každý dotaz deset vláken** spolu s anotací, jestli se **vlákno týká dotazu**. Vlákna řadíme podle podobnosti k dotazu.
  - Tyto datové sady byly použity pro evaluaci v obou následujících experimentech.

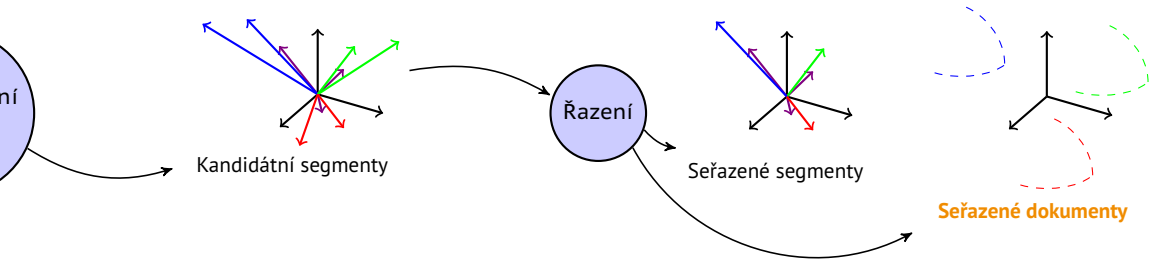
Segmentované vyhledávání

- Vyhledávač navržený v projektu Omega indexuje a navrací **tématicky koherentní segmenty dokumentů**. Vyhledávače však často musí navracet celé dokumenty.

FAKULTA INFORMATIKY
Masarykova univerzita

Segmentované vyhledávání

- Vyhledávač navržený v projektu Omega indexuje a navrací tématicky koherentní segmenty dokumentů. Vyhledávače však často musí navracet celé dokumenty.
- V rámci experimentu jsem vyhledávač rozšířil o komponentu, která **agreguje podobnost segmentů vůči dotazu** do odhadu podobnosti dokumentu vůči dotazu:



FAKULTA
INFORMATIKY
Masarykova univerzita

Segmentované vyhledávání

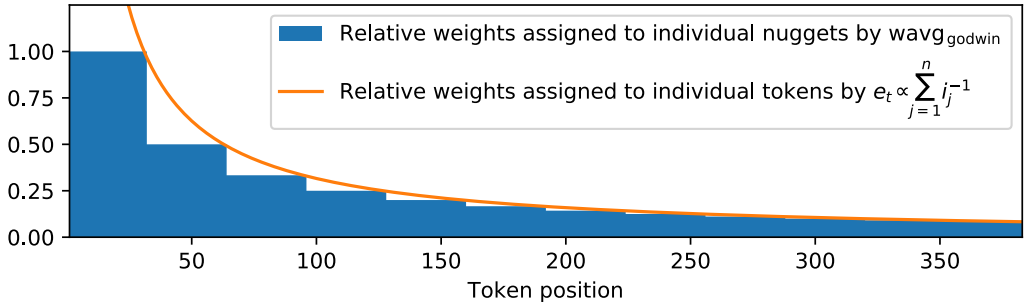
- Vyhledávač navržený v projektu Omega indexuje a navrací tématicky koherentní segmenty dokumentů. Vyhledávače však často musí navracet celé dokumenty.
- V rámci experimentu jsem vyhledávač rozšířil o komponentu, která agreguje podobnost segmentů vůči dotazu do odhadu podobnosti dokumentu vůči dotazu.
  - **Vlákna představují dokumenty**, **otázka a komentáře uvnitř vláken představují segmenty**.

Segmentované vyhledávání

- Vyhledávač navržený v projektu Omega indexuje a navrací tématicky koherentní segmenty dokumentů. Vyhledávače však často musí navracet celé dokumenty.
- V rámci experimentu jsem vyhledávač rozšířil o komponentu, která agreguje podobnost segmentů vůči dotazu do odhadu podobnosti dokumentu vůči dotazu.
  - Vlákna představují dokumenty, otázka a komentáře uvnitř vláken představují segmenty.
  - S ohledem na analýzu datových sad podúlohy 3a je hlavním agregačním mechanismem **vážený průměr s vahou i^{-1} pro komentář na pozici i**. Tento mechanismus **porazil vítěze** ročníků 2016 a 2017 soutěže SemEval.

Segmentované vyhledávání

- Vyhledávač navržený v projektu Omega indexuje a navrací tématicky koherentní segmenty dokumentů. Vyhledávače však často musí navracet celé dokumenty.
- V rámci experimentu jsem vyhledávač rozšířil o komponentu, která agreguje podobnost segmentů vůči dotazu do odhadu podobnosti dokumentu vůči dotazu.
  - Vlákna představují dokumenty, otázka a komentáře uvnitř vláken představují segmenty.
  - S ohledem na analýzu datových sad podúlohy 3a je hlavním agregačním mechanismem **vážený průměr s vahou i^{-1} pro komentář na pozici i**. Tento mechanismus porazil vítěze ročníků 2016 a 2017 soutěže SemEval.
  - Pro srovnání byl otestován i **vyhledávač bez segmentace**, který analogickým způsobem **váží jednotlivá slova dokumentu**. Tento vyhledávač **byl poražen baseline výsledkem**.



Obrázek: Poměrný dopad jednotlivých slov v dokumentu na výsledný odhad podobnosti při váženém průměru jednotlivých segmentů (modře vyplněný graf) a při váženém průměru jednotlivých slov (oranžový graf).

Segmentované vyhledávání

- Vyhledávač navržený v projektu Omega indexuje a navrací tématicky koherentní segmenty dokumentů. Vyhledávače však často musí navracet celé dokumenty.
- V rámci experimentu jsem vyhledávač rozšířil o komponentu, která agreguje podobnost segmentů vůči dotazu do odhadu podobnosti dokumentu vůči dotazu.
  - Vlákna představují dokumenty, otázka a komentáře uvnitř vláken představují segmenty.
  - S ohledem na analýzu datových sad podúlohy 3a je hlavním agregačním mechanismem **vážený průměr s vahou i^{-1} pro komentář na pozici i**. Tento mechanismus porazil vítěze ročníků 2016 a 2017 soutěže SemEval.
  - Pro srovnání byl otestován i **vyhledávač bez segmentace**, který analogickým způsobem **váží jednotlivá slova dokumentu**. Tento vyhledávač **byl poražen baseline výsledkem**.

Segmentované vyhledávání

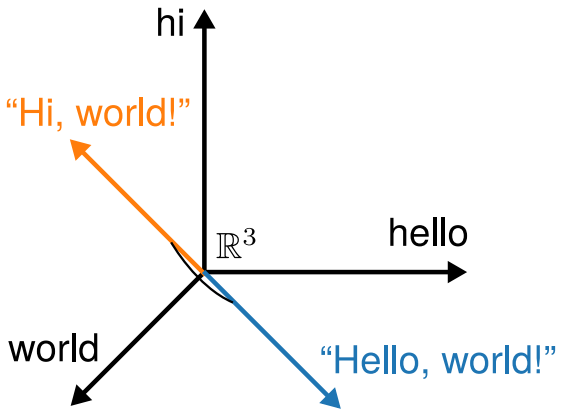
- Vyhledávač navržený v projektu Omega indexuje a navrací tématicky koherentní segmenty dokumentů. Vyhledávače však často musí navracet celé dokumenty.
- V rámci experimentu jsem vyhledávač rozšířil o komponentu, která agreguje podobnost segmentů vůči dotazu do odhadu podobnosti dokumentu vůči dotazu.
  - Vlákna představují dokumenty, otázka a komentáře uvnitř vláken představují segmenty.
  - S ohledem na analýzu datových sad podúlohy 3a je hlavním agregačním mechanismem vážený průměr s vahou i^{-1} pro komentář na pozici i. Tento mechanismus porazil vítěze ročníků 2016 a 2017 soutěže SemEval.
  - Pro srovnání byl otestován i vyhledávač bez segmentace, který analogickým způsobem váží jednotlivá slova dokumentu. Tento vyhledávač byl poražen baseline výsledkem.
  - Pro srovnání byl otestován i **vyhledávač bez segmentace**, který **z vlákna zachovává pouze úvodní otázku**. Tento vyhledávač **porazil baseline výsledek**, **ale ne vítěze soutěže**.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí **slovních histogramů (bag of words)**.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- **Podobnost** dvou dokumentů **je dána kosinem úhlu** mezi histogramy; standardní model předpokládá, že **histogramy zadávají souřadnice v ortogonální bázi**.



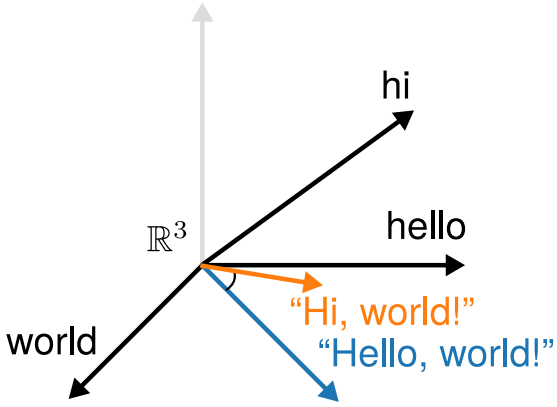
Obrázek: Standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- **Podobnost** dvou dokumentů **je dána kosinem úhlu** mezi histogramy; standardní model předpokládá, že **histogramy zadávají souřadnice v ortogonální bázi**.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- Podobnost dvou dokumentů je dána kosinem úhlu mezi histogramy; standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.
- Vyhledávač jsem rozšířil, aby **nepředpokládal, že bázové vektory jsou ortogonální**.



Obrázek: Rozšířený model předpokládá, že histogramy zadávají souřadnice v libovolné bázi.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- Podobnost dvou dokumentů je dána kosinem úhlu mezi histogramy; standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.
- Vyhledávač jsem rozšířil, aby **nepředpokládal, že bázové vektory jsou ortogonální**.

FAKULTA INFORMATIKY
Masarykova univerzita

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- Podobnost dvou dokumentů je dána kosinem úhlu mezi histogramy; standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.
- Vyhledávač jsem rozšířil, aby nepředpokládal, že bázové vektory jsou ortogonální.
  - **Skalární součin** bázových vektorů je **zadán Gramovou maticí S** velikosti n.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- Podobnost dvou dokumentů je dána kosinem úhlu mezi histogramy; standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.
- Vyhledávač jsem rozšířil, aby nepředpokládal, že bázové vektory jsou ortogonální.
  - Skalární součin bázových vektorů je zadán Gramovou maticí **S** velikosti *n*.
  - Popsal jsem, kdy lze **vypočítat kosinus úhlu v konstatním čase**.

FAKULTA INFORMATIKY
Masarykova univerzita

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- Podobnost dvou dokumentů je dána kosinem úhlu mezi histogramy; standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.
- Vyhledávač jsem rozšířil, aby nepředpokládal, že bázové vektory jsou ortogonální.
  - Skalární součin bázových vektorů je zadán Gramovou maticí **S** velikosti n.
  - Popsal jsem, kdy lze vypočítat kosinus úhlu v konstatním čase.
  - Popsal jsem, jak **v čase $\mathcal{O}(n^3)$ vypočítat matici přechodu do ortogonální báze**.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- Podobnost dvou dokumentů je dána kosinem úhlu mezi histogramy; standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.
- Vyhledávač jsem rozšířil, aby nepředpokládal, že bázové vektory jsou ortogonální.
  - Skalární součin bázových vektorů je zadán Gramovou maticí **S** velikosti n.
  - Popsal jsem, kdy lze vypočítat kosinus úhlu v konstatním čase.
  - Popsal jsem, jak v čase $\mathcal{O}(n^3)$ vypočítat matici přechodu do ortogonální báze.
  - Zadefinoval a evaluoval jsem trojici matic **S**, které **modelují různé rysy synonymie**.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- Podobnost dvou dokumentů je dána kosinem úhlu mezi histogramy; standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.
- Vyhledávač jsem rozšířil, aby nepředpokládal, že bázové vektory jsou ortogonální.
  - Skalární součin bázových vektorů je zadán Gramovou maticí **S** velikosti n.
  - Popsal jsem, kdy lze vypočítat kosinus úhlu v konstatním čase.
  - Popsal jsem, jak v čase $\mathcal{O}(n^3)$ vypočítat matici přechodu do ortogonální báze.
  - Zadefinoval a evaluoval jsem trojici matic **S**, které modelují různé rysy synonymie.
  - Diskutoval jsem **implementaci ve vektorových databázích a invertovaných indexech**.

Modelování synonymie

- Vyhledávač navržený v projektu Omega reprezentuje dokumenty pomocí slovních histogramů (bag of words).
- Podobnost dvou dokumentů je dána kosinem úhlu mezi histogramy; standardní model předpokládá, že histogramy zadávají souřadnice v ortogonální bázi.
- Vyhledávač jsem rozšířil, aby nepředpokládal, že bázové vektory jsou ortogonální.
  - Skalární součin bázových vektorů je zadán Gramovou maticí **S** velikosti n.
  - Popsal jsem, kdy lze vypočítat kosinus úhlu v konstatním čase.
  - Popsal jsem, jak v čase $\mathcal{O}(n^3)$ vypočítat matici přechodu do ortogonální báze.
  - Zadefinoval a evaluoval jsem trojici matic **S**, které modelují různé rysy synonymie.
  - Diskutoval jsem implementaci ve vektorových databázích a invertovaných indexech.
  - Dosáhl jsem **srovnatelných výsledků s vítězi** ročníků 2016 a 2017 soutěže SemEval.

FAKULTA INFORMATIKY
Masarykova univerzita

Závěr a budoucí výzkum

- Objevil jsem **statisticky významný vztah mezi pozicí příspěvku** v diskuzi **a jeho relevancí** k tématu diskuze na datových sadách soutěže SemEval.

FAKULTA INFORMATIKY
Masarykova univerzita

Závěr a budoucí výzkum

- Objevil jsem **statisticky významný vztah mezi pozicí příspěvku** v diskuzi **a jeho relevancí** k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na **nezávislých datových sadách**.

FAKULTA INFORMATIKY
Masarykova univerzita

Závěr a budoucí výzkum

- Objevil jsem statisticky významný vztah mezi pozicí příspěvku v diskuzi a jeho relevancí k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na nezávislých datových sadách.
- Popsal jsem **dvojici technik**, pomocí kterých lze **zlepšit kvalitu výsledků** běžného vyhledávače **na úroveň *state-of-the-art*** výsledků ze soutěže SemEval.

FAKULTA INFORMATIKY
Masarykova univerzita

Závěr a budoucí výzkum

- Objevil jsem statisticky významný vztah mezi pozicí příspěvku v diskuzi a jeho relevancí k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na nezávislých datových sadách.
- Popsal jsem **dvojici technik**, pomocí kterých lze **zlepšit kvalitu výsledků** běžného vyhledávače **na úroveň *state-of-the-art*** výsledků ze soutěže SemEval.
  - Budoucí výzkum by se měl zaměřit na evaluaci systému, který **implementuje obě techniky současně**, ideálně na nezávislých datových sadách.

FAKULTA INFORMATIKY
Masarykova univerzita

Závěr a budoucí výzkum

- Objevil jsem statisticky významný vztah mezi pozicí příspěvku v diskuzi a jeho relevancí k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na nezávislých datových sadách.
- Popsal jsem dvojici technik, pomocí kterých lze zlepšit kvalitu výsledků běžného vyhledávače na úroveň *state-of-the-art* výsledků ze soutěže SemEval.
  - Budoucí výzkum by se měl zaměřit na evaluaci systému, který implementuje obě techniky současně, ideálně na nezávislých datových sadách.
- Kapitolu o segmentaci jsem nezávisle zaslal na konferenci ECIR 2018.

FAKULTA INFORMATIKY
Masarykova univerzita

Závěr a budoucí výzkum

- Objevil jsem statisticky významný vztah mezi pozicí příspěvku v diskuzi a jeho relevancí k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na nezávislých datových sadách.
- Popsal jsem dvojici technik, pomocí kterých lze zlepšit kvalitu výsledků běžného vyhledávače na úroveň *state-of-the-art* výsledků ze soutěže SemEval.
  - Budoucí výzkum by se měl zaměřit na evaluaci systému, který implementuje obě techniky současně, ideálně na nezávislých datových sadách.
- Kapitolu o segmentaci jsem nezávisle zaslal na konferenci ECIR 2018.
  - Jeden z recenzentů **navrhl článek na *best paper award***.

Závěr a budoucí výzkum

- Objevil jsem statisticky významný vztah mezi pozicí příspěvku v diskuzi a jeho relevancí k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na nezávislých datových sadách.
- Popsal jsem dvojici technik, pomocí kterých lze zlepšit kvalitu výsledků běžného vyhledávače na úroveň *state-of-the-art* výsledků ze soutěže SemEval.
  - Budoucí výzkum by se měl zaměřit na evaluaci systému, který implementuje obě techniky současně, ideálně na nezávislých datových sadách.
- Kapitolu o segmentaci jsem nezávisle zaslal na konferenci ECIR 2018.
  - Jeden z recenzentů navrhl článek na *best paper award*.
  - Článek byl **zamítnut** kvůli údajné **nedostatečné obecnosti použitých datových sad**.

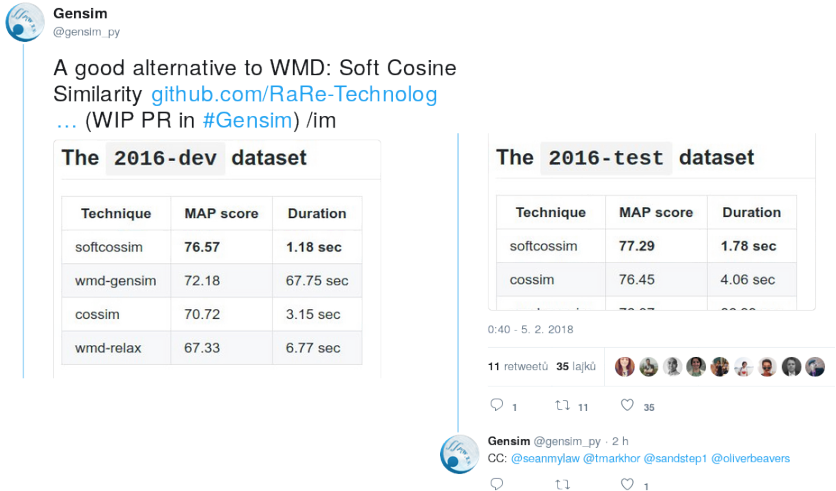
Závěr a budoucí výzkum

- Objevil jsem statisticky významný vztah mezi pozicí příspěvku v diskuzi a jeho relevancí k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na nezávislých datových sadách.
- Popsal jsem dvojici technik, pomocí kterých lze zlepšit kvalitu výsledků běžného vyhledávače na úroveň *state-of-the-art* výsledků ze soutěže SemEval.
  - Budoucí výzkum by se měl zaměřit na evaluaci systému, který implementuje obě techniky současně, ideálně na nezávislých datových sadách.
- Kapitolu o segmentaci jsem nezávisle zaslal na konferenci ECIR 2018.
  - Jeden z recenzentů navrhl článek na *best paper award*.
  - Článek byl zamítnut kvůli údajné nedostatečné obecnosti použitých datových sad.
- Načrtnul jsem, jak by mohl agregační mechanismus **využívat strojové učení**.

Závěr a budoucí výzkum

- Objevil jsem statisticky významný vztah mezi pozicí příspěvku v diskuzi a jeho relevancí k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na nezávislých datových sadách.
- Popsal jsem dvojici technik, pomocí kterých lze zlepšit kvalitu výsledků běžného vyhledávače na úroveň *state-of-the-art* výsledků ze soutěže SemEval.
  - Budoucí výzkum by se měl zaměřit na evaluaci systému, který implementuje obě techniky současně, ideálně na nezávislých datových sadách.
- Kapitolu o segmentaci jsem nezávisle zaslal na konferenci ECIR 2018.
  - Jeden z recenzentů navrhl článek na *best paper award*.
  - Článek byl zamítnut kvůli údajné nedostatečné obecnosti použitých datových sad.
- Načrtnul jsem, jak by mohl agregační mechanismus využívat strojové učení.
- Neortogonální model jsem **zanesl do knihovny Gensim pro modelování jazyka**.

FAKULTA INFORMATIKY
Masarykova univerzita



Obrázek: Neortogonální model po implementaci do knihovny Gensim 3.4.0.

Závěr a budoucí výzkum

- Objevil jsem statisticky významný vztah mezi pozicí příspěvku v diskuzi a jeho relevancí k tématu diskuze na datových sadách soutěže SemEval.
  - Budoucí výzkum by měl toto pozorování potvrdit na nezávislých datových sadách.
- Popsal jsem dvojici technik, pomocí kterých lze zlepšit kvalitu výsledků běžného vyhledávače na úroveň *state-of-the-art* výsledků ze soutěže SemEval.
  - Budoucí výzkum by se měl zaměřit na evaluaci systému, který implementuje obě techniky současně, ideálně na nezávislých datových sadách.
- Kapitolu o segmentaci jsem nezávisle zaslal na konferenci ECIR 2018.
  - Jeden z recenzentů navrhl článek na *best paper award*.
  - Článek byl zamítnut kvůli údajné nedostatečné obecnosti použitých datových sad.
- Načrtnul jsem, jak by mohl agregační mechanismus využívat strojové učení.
- Neortogonální model jsem **zanesl do knihovny Gensim pro modelování jazyka**.

Děkuji vám za pozornost.

Reakce na posudek vedoucího

Interpretujte obrázky 4.1 – 4.5 a způsoby nastavení prahu hustoty matice S_{rel}.

FAKULTA INFORMATIKY
Masarykova univerzita

Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic **S** je **matice $\mathbf{S}_{\text{rel}}$**, která **odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov**:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}} \leq \theta_3 \text{ a} \\ \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}}^{\theta_5} & \text{jinak.} \end{cases}$$

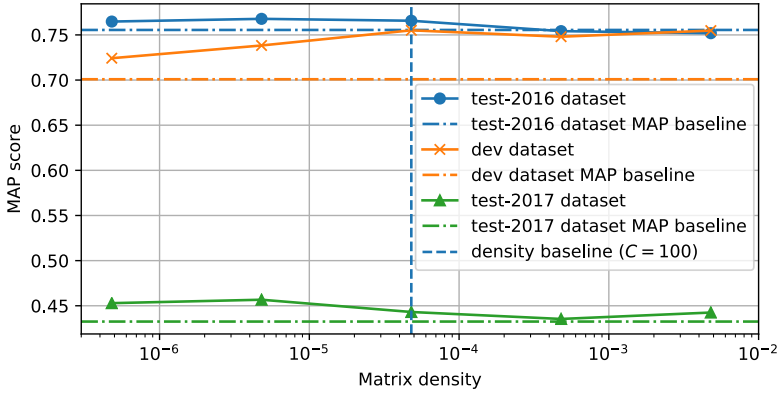
Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{rel}}$, která odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}} \leq \theta_3 \text{ a} \\ \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}}^{\theta_5} & \text{jinak.} \end{cases}$$

  - **Hustotu** matice $\mathbf{S}_{\text{rel}}$ **lze řídit parametry** C, θ_3 **a** `min_count`.



Obrázek: Graf MAP skóre a hustoty matice $\mathbf{S}_{\text{rel}}$, **při změně parametru** C od hodnoty 1 (vlevo) po hodnotu 10 000 (vpravo). Vodorovné přímky značí MAP skóre při použití kosinové podobnosti. Svislá přímka značí výchozí hodnotu parametru C.

FAKULTA INFORMATIKY
Masarykova univerzita

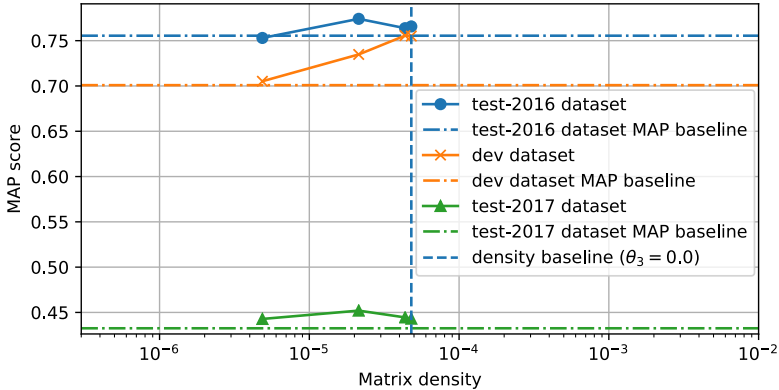
Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{rel}}$, která odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}} \leq \theta_3 \text{ a} \\ \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}}^{\theta_5} & \text{jinak.} \end{cases}$$

  - **Hustotu** matice $\mathbf{S}_{\text{rel}}$ **lze řídit parametry** C, θ_3 **a** `min_count`.



Obrázek: Graf MAP skóre a hustoty matice $\mathbf{S}_{\text{rel}}$, **při změně parametru** θ_3 od hodnoty 0,8 (vlevo) po hodnotu 0 (vpravo). Vodorovné přímky značí MAP skóre při použití kosinové podobnosti. Svislá přímka značí výchozí hodnotu parametru θ_3.

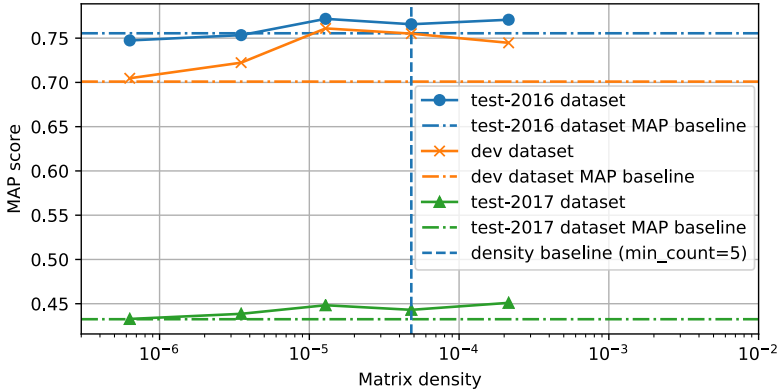
Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{rel}}$, která odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}} \leq \theta_3 \text{ a} \\ \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}}^{\theta_5} & \text{jinak.} \end{cases}$$

  - **Hustotu** matice $\mathbf{S}_{\text{rel}}$ **lze řídit parametry** C, θ_3 **a** `min_count`.



Obrázek: Graf MAP skóre a hustoty matice $\mathbf{S}_{\text{rel}}$, **při změně parametru `min_count`** od hodnoty 5 000 (vlevo) po hodnotu 0 (vpravo). Vodorovné přímky značí MAP skóre při použití kosinové podobnosti. Svislá přímka značí výchozí hodnotu parametru `min_count`.

Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{rel}}$, která odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}} \leq \theta_3 \text{ a} \\ \langle \mathbf{v}_i/\|\mathbf{v}_i\|, \mathbf{v}_j/\|\mathbf{v}_j\| \rangle_{\mathbb{X}}^{\theta_5} & \text{jinak.} \end{cases}$$

  - **Hustotu** matice $\mathbf{S}_{\text{rel}}$ **lze řídit parametry** C, θ_3 **a** `min_count`.

Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic **S** je matice $\mathbf{S}_{\text{rel}}$, která odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov.
  - Hustotu matice $\mathbf{S}_{\text{rel}}$ lze řídit parametry C, θ_3 a `min_count`.
- Druhá z popsaných a testovaných matic **S** je **matice $\mathbf{S}_{\text{lev}}$**, která **odvozuje úhel mezi dvěma bázovými vektory z Levenshteinovy vzdálenosti příslušných slov**:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \theta_1 \left(1 - \frac{\text{edit}(i,j)}{\max(b_i,b_j)}\right)^{\theta_2} \leq \theta_3 \text{ a} \\ \theta_1 \left(1 - \frac{\text{edit}(i,j)}{\max(b_i,b_j)}\right)^{\theta_2} & \text{jinak.} \end{cases}$$

FAKULTA INFORMATIKY
Masarykova univerzita

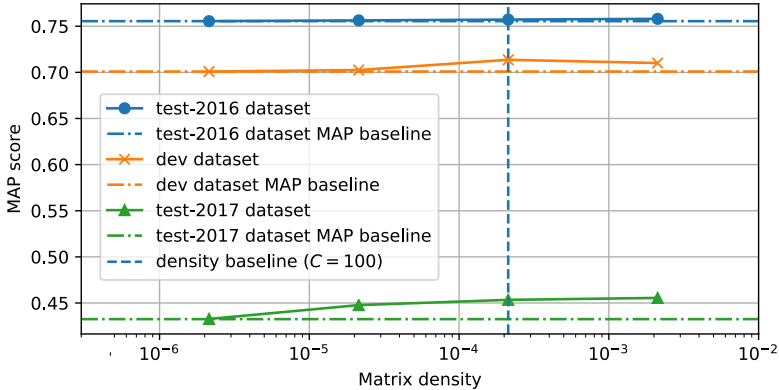
Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{rel}}$, která odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov.
  - Hustotu matice $\mathbf{S}_{\text{rel}}$ lze řídit parametry C, θ_3 a `min_count`.
- Druhá z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{lev}}$, která odvozuje úhel mezi dvěma bázovými vektory z Levenshteinovy vzdálenosti příslušných slov:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \theta_1 \left(1 - \frac{\text{edit}(i,j)}{\max(b_i,b_j)}\right)^{\theta_2} \leq \theta_3 \text{ a} \\ \theta_1 \left(1 - \frac{\text{edit}(i,j)}{\max(b_i,b_j)}\right)^{\theta_2} & \text{jinak.} \end{cases}$$

  - Hustotu matice $\mathbf{S}_{\text{lev}}$ **lze řídit parametry** C **a** θ_3.



Obrázek: Graf MAP skóre a hustoty matice $\mathbf{S}_{\text{lev}}$, **při změně parametru** C od hodnoty 1 (vlevo) po hodnotu 1 000 (vpravo). Vodorovné přímky značí MAP skóre při použití kosinové podobnosti. Svislá přímka značí výchozí hodnotu parametru C.

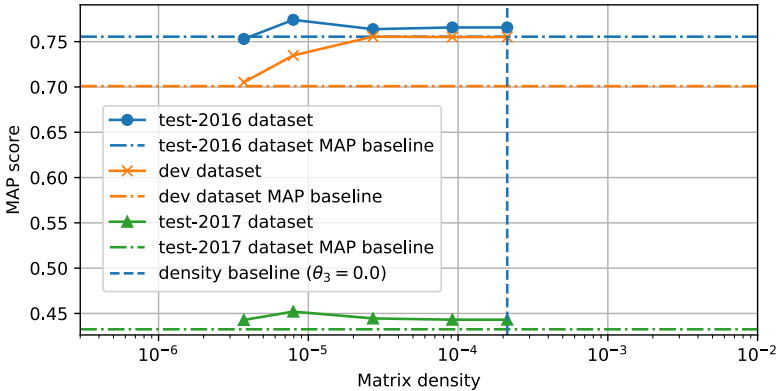
Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{rel}}$, která odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov.
  - Hustotu matice $\mathbf{S}_{\text{rel}}$ lze řídit parametry C, θ_3 a `min_count`.
- Druhá z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{lev}}$, která odvozuje úhel mezi dvěma bázovými vektory z Levenshteinovy vzdálenosti příslušných slov:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \theta_1 \left(1 - \frac{\text{edit}(i,j)}{\max(b_i,b_j)}\right)^{\theta_2} \leq \theta_3 \text{ a} \\ \theta_1 \left(1 - \frac{\text{edit}(i,j)}{\max(b_i,b_j)}\right)^{\theta_2} & \text{jinak.} \end{cases}$$

  - Hustotu matice $\mathbf{S}_{\text{lev}}$ **lze řídit parametry** C **a** θ_3.



Obrázek: Graf MAP skóre a hustoty matice $\mathbf{S}_{\text{lev}}$, **při změně parametru** θ_3 od hodnoty 0,8 (vlevo) po hodnotu 0 (vpravo). Vodorovné přímky značí MAP skóre při použití kosinové podobnosti. Svislá přímka značí výchozí hodnotu parametru θ_3.

Reakce na posudek vedoucího

Interpretujte obrázky 4.1–4.5 a způsoby nastavení prahu hustoty matice $\mathbf{S}_{\text{rel}}$.

- První z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{rel}}$, která odvozuje úhel mezi dvěma bázovými vektory z úhlu mezi embeddingy příslušných slov.
  - Hustotu matice $\mathbf{S}_{\text{rel}}$ lze řídit parametry C, θ_3 a `min_count`.
- Druhá z popsaných a testovaných matic $\mathbf{S}$ je matice $\mathbf{S}_{\text{lev}}$, která odvozuje úhel mezi dvěma bázovými vektory z Levenshteinovy vzdálenosti příslušných slov:

$$s_{ij} = \begin{cases} 1 & \text{pokud } i = j, \\ 0 & \text{pokud } \theta_1 \left(1 - \frac{\text{edit}(i,j)}{\max(b_i,b_j)}\right)^{\theta_2} \leq \theta_3 \text{ a} \\ \theta_1 \left(1 - \frac{\text{edit}(i,j)}{\max(b_i,b_j)}\right)^{\theta_2} & \text{jinak.} \end{cases}$$

  - Hustotu matice $\mathbf{S}_{\text{lev}}$ **lze řídit parametry** C **a** θ_3.

Reakce na posudek oponenta

K obhajobě práce mám na autora jednu otázku: V kapitole 4.6 je popsána ***metoda expanze dotazů****, která umožňuje nalézt i takové dokumenty, které neobsahují žádné termy z původního dotazu. Nabízí se však také* ***alternativní přístup****,* ***který by expandoval texty dokumentu*** *a indexoval je potom i pod expandovanými termy.* ***Jsou oba přístupy ekvivalentní****, nebo je některý z nich pro information retrieval výhodnější?*

Reakce na posudek oponenta

K obhajobě práce mám na autora jednu otázku: V kapitole 4.6 je popsána metoda expanze dotazů, která umožňuje nalézt i takové dokumenty, které neobsahují žádné termy z původního dotazu. Nabízí se však také alternativní přístup, který by expandoval texty dokumentu a indexoval je potom i pod expandovanými termy. Jsou oba přístupy ekvivalentní, nebo je některý z nich pro information retrieval výhodnější?

- V sekci 4.6 popisuji, jak lze **neortogonální model implementovat pomocí expanze dotazu** na straně klienta.

$d_2 =$ “I did enact Julius Caesar: I was killed i’ the Capitol”,

$d_3 =$ “Give␣unto␣Caesar Brutus␣Cassius choreographers␣Bosco Julius␣Caeser therefore␣unto␣Caesar Marcus␣Antonius Caesarion Gallic␣Wars Marcus␣Crassus Antoninus Catiline Seleucus Gaius␣Julius␣Caesar Theodoric Marcus␣Tullius␣Cicero unto␣Caesar emperor␣Nero

⋮

Benjamin Kenneth Philip Marcus Arthur Carl Fred Edward Jonathan Eric Frank Anthony William Richard Robert enact Capitol killed Ididn’t honestly myself I I my we the ’d ’m did was”.

Obrázek: Expanze dotazu na straně klienta.

Reakce na posudek oponenta

K obhajobě práce mám na autora jednu otázku: V kapitole 4.6 je popsána metoda expanze dotazů, která umožňuje nalézt i takové dokumenty, které neobsahují žádné termy z původního dotazu. Nabízí se však také alternativní přístup, který by expandoval texty dokumentu a indexoval je potom i pod expandovanými termy. Jsou oba přístupy ekvivalentní, nebo je některý z nich pro information retrieval výhodnější?

- V sekci 4.6 popisuji, jak lze **neortogonální model implementovat pomocí expanze dotazu** na straně klienta.

Reakce na posudek oponenta

K obhajobě práce mám na autora jednu otázku: V kapitole 4.6 je popsána metoda expanze dotazů, která umožňuje nalézt i takové dokumenty, které neobsahují žádné termy z původního dotazu. Nabízí se však také alternativní přístup, který by expandoval texty dokumentu a indexoval je potom i pod expandovanými termy. Jsou oba přístupy ekvivalentní, nebo je některý z nich pro information retrieval výhodnější?

- V sekci 4.6 popisuji, jak lze **neortogonální model implementovat pomocí expanze dotazu** na straně klienta.
  - Na serveru lze uchovávat **původní dokumenty**.

Reakce na posudek oponenta

K obhajobě práce mám na autora jednu otázku: V kapitole 4.6 je popsána metoda expanze dotazů, která umožňuje nalézt i takové dokumenty, které neobsahují žádné termy z původního dotazu. Nabízí se však také alternativní přístup, který by expandoval texty dokumentu a indexoval je potom i pod expandovanými termy. Jsou oba přístupy ekvivalentní, nebo je některý z nich pro information retrieval výhodnější?

- V sekci 4.6 popisuji, jak lze **neortogonální model implementovat pomocí expanze dotazu** na straně klienta.
  - Na serveru lze uchovávat původní dokumenty.
  - Na klientovi lze použít **rozličné matice S** bez změny obsahu indexu.

Reakce na posudek oponenta

K obhajobě práce mám na autora jednu otázku: V kapitole 4.6 je popsána metoda expanze dotazů, která umožňuje nalézt i takové dokumenty, které neobsahují žádné termy z původního dotazu. Nabízí se však také alternativní přístup, který by expandoval texty dokumentu a indexoval je potom i pod expandovanými termy. Jsou oba přístupy ekvivalentní, nebo je některý z nich pro information retrieval výhodnější?

- V sekci 4.6 popisuji, jak lze neortogonální model implementovat pomocí expanze dotazu na straně klienta.
  - Na serveru lze uchovávat původní dokumenty.
  - Na klientovi lze použít rozličné matice **S** bez změny obsahu indexu.
- Při expanzi indexovaných dokumentů **dochází až k n-násobnému nárůstu objemu uchovávaných dat** a **změna matic S vyžaduje opětovnou indexaci** všech dokumentů.

Děkuji vám za pozornost.