FACULTY
OF INFORMATICS
Masaryk University

Weighting of Passages in Question Answering

An Empirical Evaluation of the Godwin's Law

Vít Novotný, witiko@mail.muni.cz
Petr Sojka, sojka@fi.muni.cz

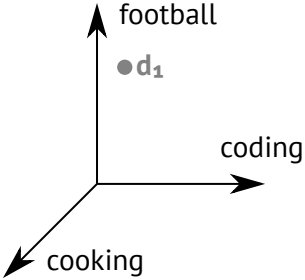
December 7, 2018

FACULTY
OF INFORMATICS
Masaryk University

Introduction

- The **vector space model** (tf-idf) is **well-understood** and **scalable**, let's keep it.

d_1 Player with 350 or more goals in a domestic league?	**football** coding cooking

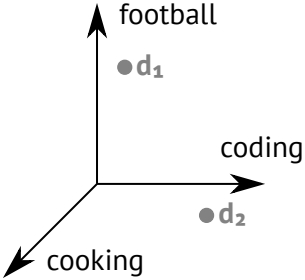


FACULTY
OF INFORMATICS
Masaryk University

Introduction

- The **vector space model** (tf-idf) is **well-understood** and **scalable**, let's keep it.

d_1 Player with 350 or more goals in a domestic league?	**football** coding cooking
d_2 Why is using pixels as a unit of measure bad in web design?	football **coding** cooking

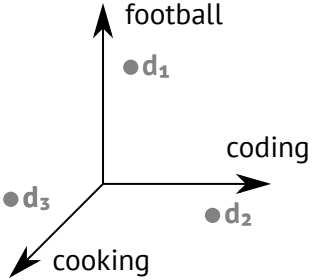


FACULTY
OF INFORMATICS
Masaryk University

Introduction

- The **vector space model** (tf-idf) is **well-understood** and **scalable**, let's keep it.

d_1 Player with 350 or more goals in a domestic league?	**football**	coding	cooking
d_2 Why is using pixels as a unit of measure bad in web design?	football	**coding**	cooking
d_3 Can I use red wine instead of white wine for chicken scallops?	football	coding	**cooking**



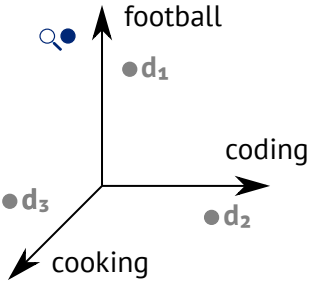
FACULTY OF INFORMATICS
Masaryk University

Introduction

- The **vector space model** (tf-idf) is **well-understood** and **scalable**, let's keep it.

d_1 Player with 350 or more goals in a domestic league?	**football**	coding	cooking
d_2 Why is using pixels as a unit of measure bad in web design?	football	**coding**	cooking
d_3 Can I use red wine instead of white wine for chicken scallops?	football	coding	**cooking**

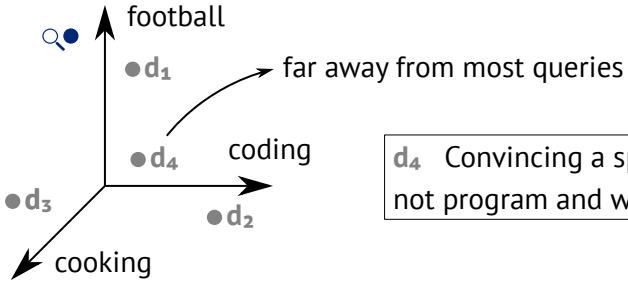
football top-scoring players



FACULTY
OF INFORMATICS
Masaryk University

Introduction

- The **vector space model** (tf-idf) is **well-understood** and **scalable**, let's keep it.
- **Problem: Long documents** that range **many topics** are almost **never retrieved**!



d_4 Convincing a spouse to cook, not program and watch football?	football coding cooking

FACULTY
OF INFORMATICS
Masaryk University

Introduction

- The **vector space model** (tf-idf) is **well-understood** and **scalable**, let's keep it.
- **Problem:** **Long documents** that range **many topics** are almost **never retrieved**!
- **Solution:** **Segmentation** to semantically coherent **passages**.

But what if we **need** to retrieve **full documents**?

FACULTY
OF INFORMATICS
Masaryk University

System

- Our system **segments** both **queries** and **indexed documents**.

u_1 I did enact Julius Caesar	v_1 So let it be with Caesar
u_2 I was killed i' the Capitol	v_2 The noble Brutus hath told you
u_3 Brutus killed me	v_3 Caesar was ambitious

FACULTY OF INFORMATICS
Masaryk University

System

- Our system **segments** both **queries** and **indexed documents**.
- At query time, we construct a **matrix $\mathbf{M}_{uv}$ of similarities** between the **segments of the query** u (rows) and the **segments of a document** v (columns).

u_1 I did enact Julius **Caesar**	v_1 So let it be with **Caesar**
u_2 I **was** killed i' **the** Capitol	v_2 **The** noble **Brutus** hath told you
u_3 **Brutus** killed me	v_3 **Caesar was** ambitious

$$\rightsquigarrow \mathbf{M}_{uv} = \begin{pmatrix} 0.18 & 0 & 0.26 \\ 0 & 0.16 & 0.24 \\ 0 & 0.24 & 0 \end{pmatrix}$$

FACULTY
OF INFORMATICS
Masaryk University

System

- Our system **segments** both **queries** and **indexed documents**.
- At query time, we construct a **matrix $\mathbf{M}_{uv}$ of similarities** between the **segments of the query** u (rows) and the **segments of a document** v (columns).
- Then, we **reduce $\mathbf{M}_{uv}$** to an **aggregate similarity**.

$$\mathbf{M}_{uv} = \begin{pmatrix} 0.18 & 0 & 0.26 \\ 0 & 0.16 & 0.24 \\ 0 & 0.24 & 0 \end{pmatrix} \quad \rightsquigarrow \quad \bigoplus_k \bigominus_l m_{kl} = 0.205$$

FACULTY
OF INFORMATICS
Masaryk University

System

- Our system **segments** both **queries** and **indexed documents**.
- At query time, we construct a **matrix $\mathbf{M}_{uv}$ of similarities** between the **segments of the query u** (rows) and the **segments of a document v** (columns).
- Then, we **reduce $\mathbf{M}_{uv}$** to an **aggregate similarity**.

The devil is in the detail. **How** exactly do we **aggregate $\mathbf{M}_{uv}$**?

FACULTY
OF INFORMATICS
Masaryk University

Datasets

- We **evaluated** our system on **SemEval** 2016 and 2017 **question answering** datasets.

FACULTY
OF INFORMATICS
Masaryk University

Datasets

- We **evaluated** our system on **SemEval** 2016 and 2017 **question answering** datasets.
- We focused on the following **two subtasks** of the SemEval question answering task:
  A) Given a **question**, **rank** the first ten **comments** by relevance.

Datasets

- We **evaluated** our system on **SemEval** 2016 and 2017 **question answering** datasets.
- We focused on the following **two subtasks** of the SemEval question answering task:
  - A) Given a **question**, **rank** the first ten **comments** by relevance.
  - B) Given an **original question**, **rank** ten **related questions** by relevance.

FACULTY
OF INFORMATICS
Masaryk University

Datasets

- We **evaluated** our system on **SemEval** 2016 and 2017 **question answering** datasets.
- We focused on the following **two subtasks** of the SemEval question answering task:
  - A) Given a **question**, **rank** the first ten **comments** by relevance.
  - B) Given an **original question**, **rank** ten **related questions** by relevance.
- **Subtask A** comes with a **training dataset** of **2,654 questions**. We **analyzed** these to learn how to **aggregate** $\mathbf{M}_{UV}$.

FACULTY
OF INFORMATICS
Masaryk University

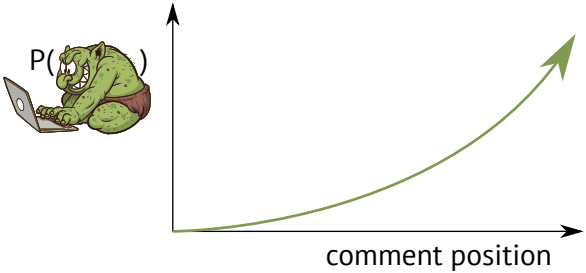
Datasets

- We **evaluated** our system on **SemEval** 2016 and 2017 **question answering** datasets.
- We focused on the following **two subtasks** of the SemEval question answering task:
  - A) Given a **question**, **rank** the first ten **comments** by relevance.
  - B) Given an **original question**, **rank** ten **related questions** by relevance.
- **Subtask A** comes with a **training dataset** of **2,654 questions**. We **analyzed** these to learn how to **aggregate** M_{UV}.
- **Subtask B** comes with a **validation dataset** of **50 original questions** and **244 related questions** (2016), and **88 original questions** and **239 related questions** (2017). We used these to **evaluate** our system. Question text and comments are the **segments**.

FACULTY
OF INFORMATICS
Masaryk University

Analysis

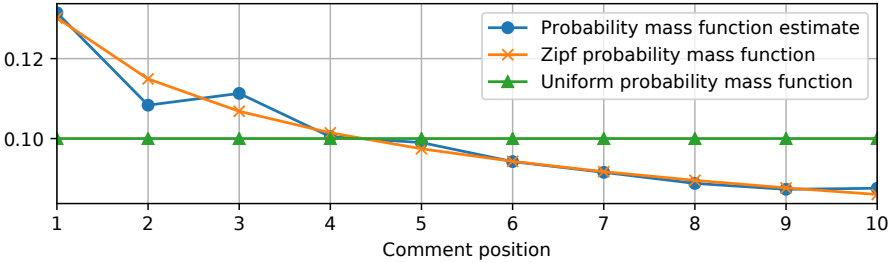
- In 1991, **Mike Godwin** posited that **later comments** are likely to be **less relevant**.



FACULTY
OF INFORMATICS
Masaryk University

Analysis

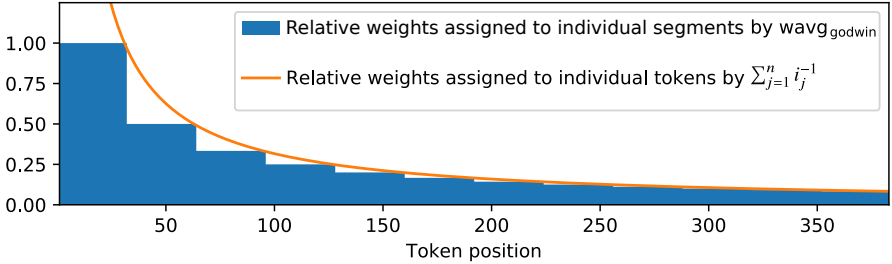
- In 1991, **Mike Godwin** posited that **later comments** are likely to be **less relevant**.
- **Godwin's rule applies** to the **subtask A** dataset with **statistical significance**.



FACULTY OF INFORMATICS
Masaryk University

Analysis

- In 1991, **Mike Godwin** posited that **later comments** are likely to be **less relevant**.
- **Godwin's rule applies** to the **subtask A** dataset with **statistical significance**.
- We **aggregate M_{uv}** using **weighted average**, where the weight of a segment is proportional to the **inverse of the segment's position**. We also evaluated **weighting tokens** in **unsegmented documents**.



FACULTY
OF INFORMATICS
Masaryk University

Results

- Our system (primary) is **on-par with winners** of SemEval 2016 and 2017.

Configuration	Segm.	Text summ.	S. f. S	Aggregate s. f. S'	MAP
Primary	**Yes**		**`bfx.tfx`**	$\oplus = \mathbf{wavg}_{\mathbf{length}}$, $\ominus = \mathbf{wavg}_{\mathbf{Godwin}}$	**76.77**
SemEval-2016 task 3 subtask B winner (UH-PRHLT-primary)					***76.70***
Third contrastive	No	FirstTwoPara	`bfx.tfx`		75.21
SemEval-2016 task 3 subtask B IR baseline					*74.75*
First contrastive	No		`bfx.tfx`		73.94
Second contrastive	No		`bfx.tfx`, Godwin		70.28

FACULTY
OF INFORMATICS
Masaryk University

Results

- Our system (primary) is **on-par with winners** of SemEval 2016 and 2017.

Configuration	Segm.	Text summ.	S. f. S	Aggregate s. f. S'	MAP
Primary	**Yes**		**`bfx.tfx`**	$\oplus = \mathbf{wavg}_{\mathbf{length}}$, $\ominus = \mathbf{wavg}_{\mathbf{Godwin}}$	**47.45**
SemEval-2017 task 3 subtask B winner (SimBow-primary)					***47.22***
Third contrastive	No	FirstTwoPara	`bfx.tfx`		44.67
SemEval-2017 task 3 subtask B IR baseline					*41.85*
Second contrastive	No		`bfx.tfx`, Godwin		37.18
First contrastive	No		`bfx.tfx`		36.82

FACULTY
OF INFORMATICS
Masaryk University

Results

- Our system (primary) is **on-par with winners** of SemEval 2016 and 2017.
- **Weighting tokens** (second contrastive) does **worse than baseline**, which shows that **segments** are the **correct level of analysis**.

Configuration	Segm.	Text summ.	S. f. S	Aggregate s. f. S'	MAP
Primary	**Yes**		**`bfx.tfx`**	$\oplus = \mathbf{wavg}_{\mathbf{length}}$, $\ominus = \mathbf{wavg}_{\mathbf{Godwin}}$	**47.45**
SemEval-2017 task 3 subtask B winner (SimBow-primary)					*47.22*
Third contrastive	No	FirstTwoPara	`bfx.tfx`		44.67
SemEval-2017 task 3 subtask B IR baseline					***41.85***
Second contrastive	**No**		**`bfx.tfx`, Godwin**		**37.18**
First contrastive	No		`bfx.tfx`		36.82

FACULTY
OF INFORMATICS
Masaryk University

Conclusion

We have accomplished the following:

- We showed that **Godwin's rule applies** in question answering.

FACULTY
OF INFORMATICS
Masaryk University

Conclusion

We have accomplished the following:

- We showed that **Godwin's rule applies** in question answering.
- We achieved **state-of-the-art results** on a SemEval task with **no semantic modeling**.

FACULTY
OF INFORMATICS
Masaryk University

Conclusion

We have accomplished the following:

- We showed that **Godwin's rule applies** in question answering.
- We achieved **state-of-the-art results** on a SemEval task with **no semantic modeling**.

In our **future work**, we will focus on the following:

- We will test the hypothesis on **other datasets**.

FACULTY
OF INFORMATICS
Masaryk University

Conclusion

We have accomplished the following:

- We showed that **Godwin's rule applies** in question answering.
- We achieved **state-of-the-art results** on a SemEval task with **no semantic modeling**.

In our **future work**, we will focus on the following:

- We will test the hypothesis on **other datasets**.
- We will investigate how this generalizes **outside question answering**.

Thank you for your attention!