



GOTHENBURG, SWEDEN

SEPTEMBER 9-13, 2019



TU/e

Advances in Content-based RecSys

Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci,
Giovanni Semeraro

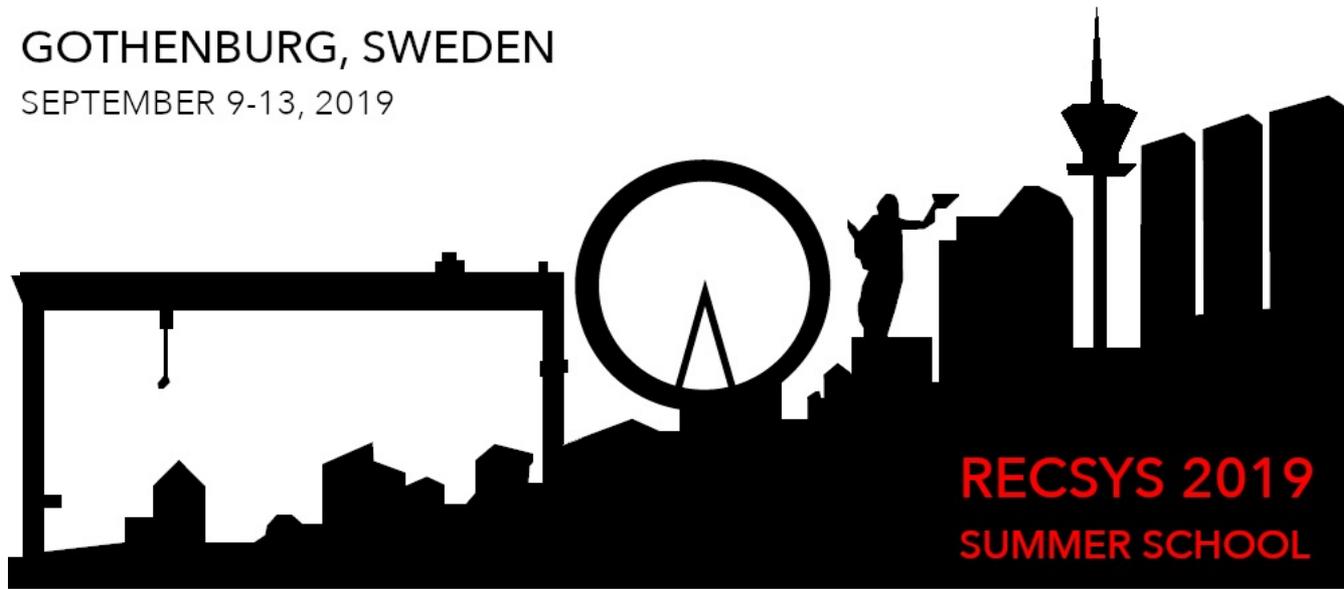
Department of Computer Science
University of Bari Aldo Moro, Italy





GOTHENBURG, SWEDEN

SEPTEMBER 9-13, 2019



TU/e

Advances in Content-based RecSys

Introduction

Giovanni Semeraro

Department of Computer Science
University of Bari Aldo Moro, Italy



About us



marco.degemmis@uniba.it



pasquale.lops@uniba.it



cataldo.musto@uniba.it



fedelucio.narducci@uniba.it



giovanni.semeraro@uniba.it



Semantic
Web
Access and
Personalization
"Antonio Bello" research group
<http://www.di.uniba.it/~swap>



in this talk...

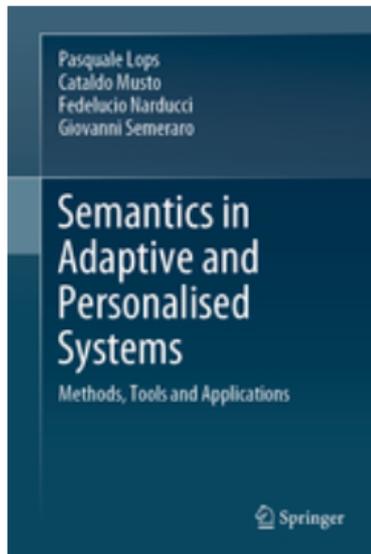
how to represent content

to improve **information access** and build a
new generation of

semantics-aware services for

user modeling &

recommender systems?



© 2019

Semantics in Adaptive and Personalised Systems

Methods, Tools and Applications

Authors: **Lops, P., Musto, C., Narducci, F., Semeraro, G.**

<https://www.springer.com/gb/book/9783030056179>

Agenda

Why?

Why do we need **content**?
Why do we need **semantics**?

How?

How to **introduce semantics**?
Basics of **Natural Language Processing**
Encoding **exogenous semantics**, i.e. *explicit* semantics
Encoding **endogenous semantics**, i.e. *implicit* semantics

What?

Explanation of Recommendations
Conversational Recommender Systems (hands-on)

Agenda

Why?

Why do we need **content**?

Why do we need **semantics**?

How?

How to **introduce semantics**?

Basics of **Natural Language Processing**

Encoding **exogenous semantics**, i.e. *explicit* semantics

Encoding **endogenous semantics**, i.e. *implicit* semantics

What?

Explanation of Recommendations

Conversational Recommender Systems (hands-on)

Why do we need content?



| Search engines need content



Cerca con Google

Mi sento fortunato

Trivial: search engines can't work without content

Why do we need content?

| |  |  |  |  |  |
|---|---|---|---|---|---|
|  | ✓ | ✓ | | ✓ | |
|  | | ✓ | | | |
|  | ✓ | | ✓ | | |
|  | | | | ✓ | ✓ |
|  | ✓ | ✓ | | | |

Recommender Systems: not trivial!

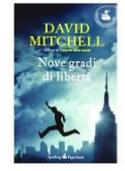
Why do we need content?

| |  |  |  |  |  |
|---|---|---|---|---|---|
|  | ✓ | ✓ | | ✓ | |
|  | | ✓ | | | |
|  | ✓ | | ✓ | | |
|  | | | | ✓ | ✓ |
|  | ✓ | ✓ | | | |

Recommender Systems can work without content

Why do we need content?

Customers Who Bought This Item Also Bought



Nove gradi di libertà
David Mitchell
Perfect Paperback
£10.45 Prime



Le ore invisibili
David Mitchell
Hardcover



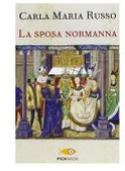
Cloud Atlas [DVD] [2013]
Tom Hanks
★★★★☆ 790
DVD
£4.99 Prime



Sogno numero 9
David Mitchell
Perfect Paperback
£10.43 Prime



Il dono della terapia
Paperback



La sposa normanna
Carla M. Russo
Perfect Paperback
£9.46 Prime



Storia Della Bellezza
Umberto Eco
Hardcover
£42.50 Prime



Puoi guarire la tua vita.
Pensa in positivo per ritrovare il benessere fisico e la serenità interiore
Louise L. Hay
Paperback

several Recommender Systems perfectly work **using no content!**

Collaborative Filtering (CF), Matrix Factorization (MF) and Tensor Factorization (TF) are state-of-the-art techniques for implementing Recommender Systems

Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata

István Pilászy^{*}
Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
pila@mit.bme.hu

Domonkos Tikk^{*,†}
Dept. of Telecom. and Media Informatics
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
tikk@mit.bme.hu

ABSTRACT

The Netflix Prize (NP) competition gave much attention to collaborative filtering (CF) approaches. Matrix factorization (MF) based CF approaches assign low dimensional feature vectors to users and items. We link CF and content-based filtering (CBF) by finding a linear transformation that transforms user or item descriptions so that they are as close as possible to the feature vectors generated by MF for CF. We propose methods for explicit feedback that are able to handle 140,000 features when feature vectors are very sparse. With movie metadata collected for the NP movies we show that the prediction performance of the methods is comparable to that of CF, and can be used to predict user preferences on new movies. We also investigate the value of movie metadata compared to movie ratings in regards of predictive power. We compare

1. INTRODUCTION

The goal of recommender systems is to give personalized recommendation on items to users. Typically the recommendation is based on the former and current activity of the users, and metadata about users and items, if available. There are two basic strategies that can be applied when generating recommendations. Collaborative filtering (CF) methods are based only on the activity of users, while content-based filtering (CBF) methods use only metadata. In this paper we propose hybrid methods, which try to benefit from both information sources. The two most important families of CF methods are matrix factorization (MF) and neighbor-based approaches. Usually, the goal of MF is to find a low dimensional representation for both users and movies, i.e. each user and movie is associated with a feature vector. Movie metadata (which

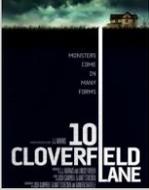
ACM RecSys 2009 paper by Netflix Challenge winners

Why do we need content?



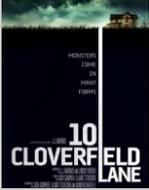
Content **can tackle some issues** of **Collaborative Filtering**

Why do we need content?

| |  |  |  |  |  |
|---|---|---|---|---|--|
|  | | | |  | |
|  | |  | | | |
|  | | |  | | |
|  | | | | |  |
|  |  | | | | |

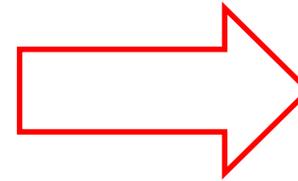
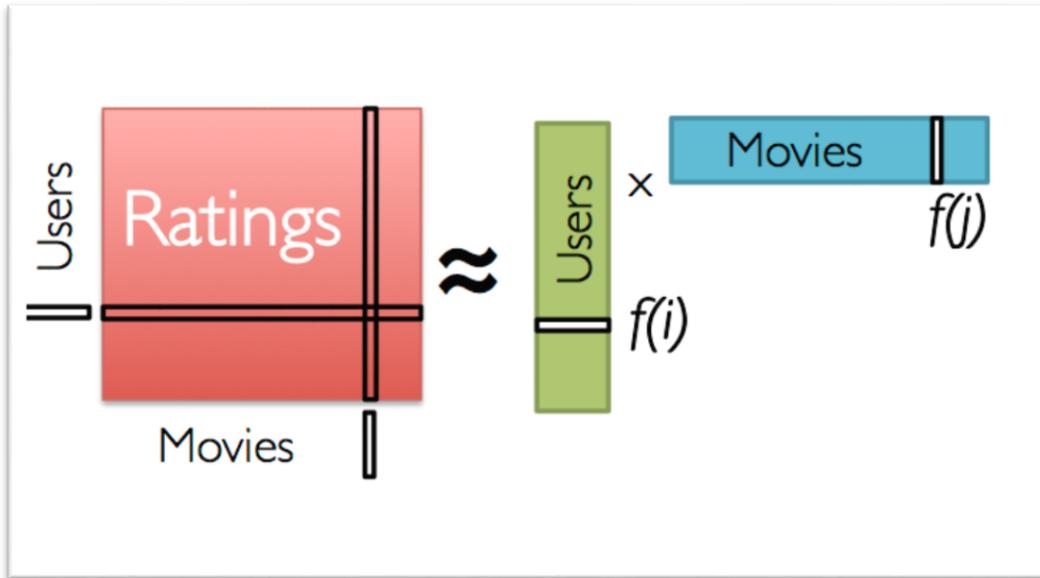
Collaborative Filtering issues: sparsity

Why do we need content?

| | | | | | |
|---|---|---|---|---|---|
| |  |  |  |  |  |
|  | ✓ | ✓ | | ✓ | |
|  | | ✓ | | | ? |
|  | | | ✓ | | ? |
|  | | | | ✓ | |
|  | ✓ | ✓ | | | |

Collaborative Filtering issues: new item problem

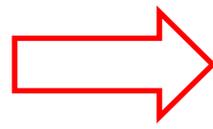
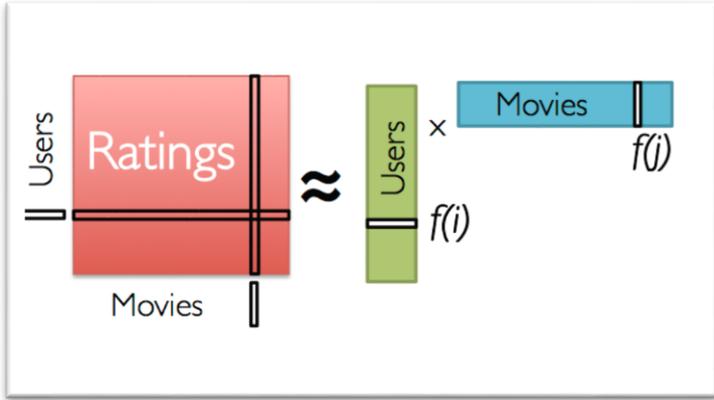
Why do we need content?



Why?

Collaborative Filtering issues: lack of transparency!

Why do we need content?



Customers Who Bought This Item Also Bought

| | | | | | | | |
|---|---|--|--|--|--|--|--|
| <p>Nove gradi di libertà David Mitchell Perfect Paperback £10.45 ✓Prime</p> | <p>Le ore invisibili David Mitchell Hardcover</p> | <p>Cloud Atlas [DVD] [2013] Tom Hanks ★★★★☆ 790 DVD £4.99 ✓Prime</p> | <p>Sogno numero 9 David Mitchell Perfect Paperback £10.43 ✓Prime</p> | <p>Il dono della terapia Paperback</p> | <p>La sposa normanna Carla M. Russo Perfect Paperback £9.46 ✓Prime</p> | <p>Storia Della Bellezza Umberto Eco Hardcover £42.50 ✓Prime</p> | <p>Puoi guarire la tua vita. Pensa in positivo per ritrovare il benessere fisico e la serenità interiore Louise L. Hay Paperback</p> |
|---|---|--|--|--|--|--|--|

Who knows the «Customers Who Bought This Item ...»?
Information Asymmetry

Collaborative Filtering issues: poor explanations!

Recap #1



Why do we need content?

- Because **search engines can't simply work** without content 😊
- **In general:** to extend and improve user modeling
- To overcome **typical issues** of **collaborative filtering**

Agenda

Why?

Why do we need **content**?

Why do we need **semantics**?

How?

How to **introduce semantics**?

Basics of **Natural Language Processing**

Encoding **exogenous semantics**, i.e. *explicit* semantics

Encoding **endogenous semantics**, i.e. *implicit* semantics

What?

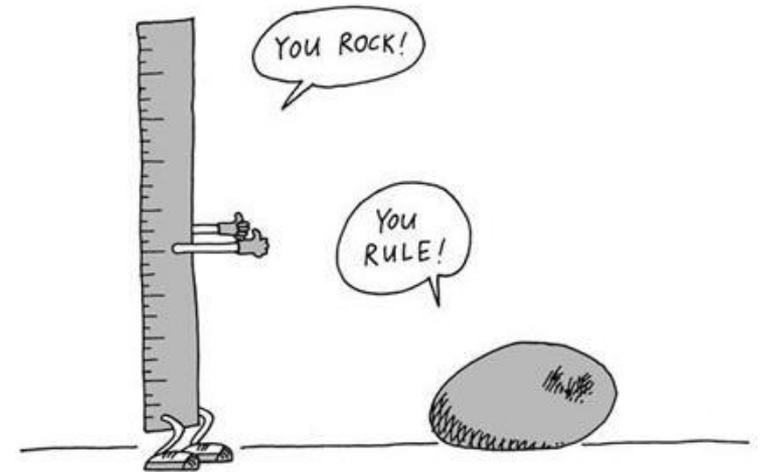
Explanation of Recommendations

Conversational Recommender Systems (hands-on)

Why do we need semantics?

Knowledge is Power hypothesis

...the power of an intelligent program to perform its task well depends primarily on the quantity and quality of knowledge it has about the task (Buchanan & Feigenbaum 1982)



Deep Rationality requires a **deep comprehension** of the information conveyed by **textual content**. To achieve that goal it is crucial to **improve the quality of user profiles** and the **effectiveness of recommender systems & intelligent information access platforms**.

Basics: Content-based RecSys (CBRS)

Suggest items similar to those the user liked in the past

Preferences-are-stable-over-time assumption

Recommendations generated by matching the **description of items** with the **profile of the user's interests**

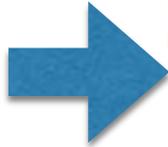
use of specific **features**



P. Lops, M. de Gemmis, and G. Semeraro. Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor (Eds.), Recommender Systems Handbook, Springer, 73-105, 2011.

Pazzani, M. J., & Billsus, D. Content-Based Recommendation Systems. The Adaptive Web. Lecture Notes in Computer Science vol. 4321, 325-341, 2007.

Basics: Content-based RecSys (CBRS)



Recommendations are generated by matching the **features stored** in the **user profile** with those describing **the items** to be recommended.

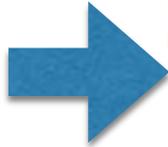


user profile



items

Basics: Content-based RecSys (CBRS)



Recommendations are generated by matching the **features stored** in the user profile with those describing the items to be recommended.



user profile



items

Lack of Semantics in User Models



“I love turkey. It’s my choice for these #holidays!”

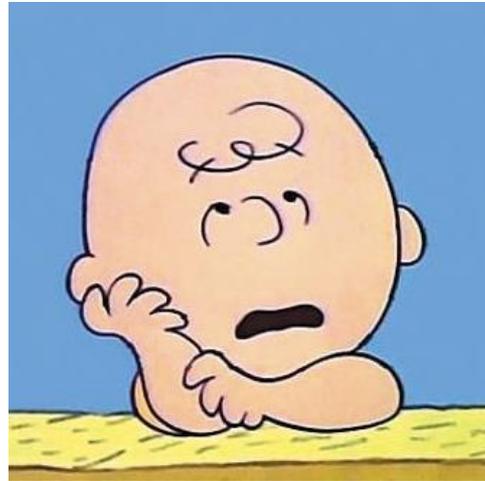


Social Media can be helpful to avoid cold start

Lack of Semantics in User Models



“I love turkey. It’s my choice for these #holidays!”



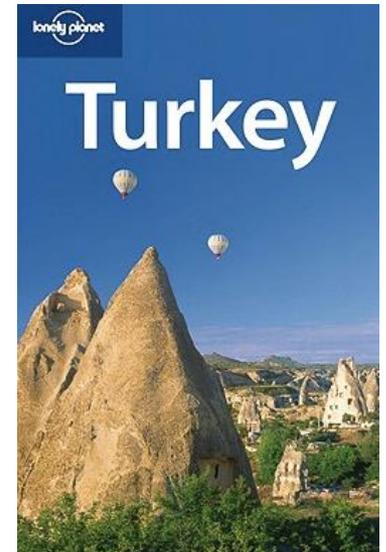
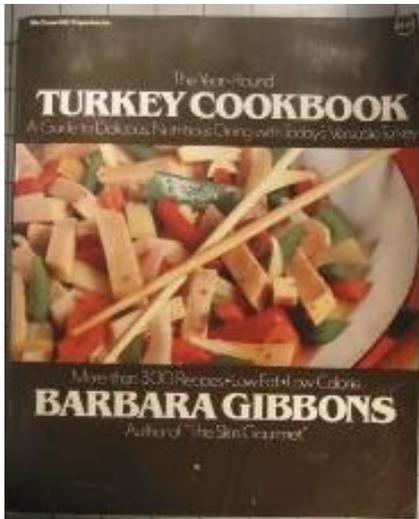
..but pure content-based representations

can't handle polysemy

Lack of Semantics in User Models



“I love turkey. It’s my choice for these #holidays!”



Pure content-based representations can easily drive a recommender system towards failures!

Lack of Semantics in User Models

...is not only about polysemy

doc1
AI is a branch of
computer science

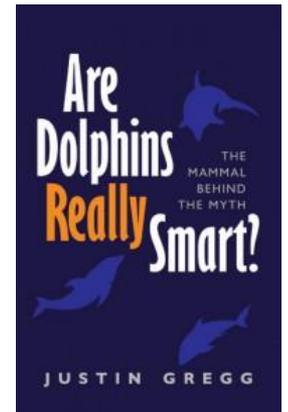
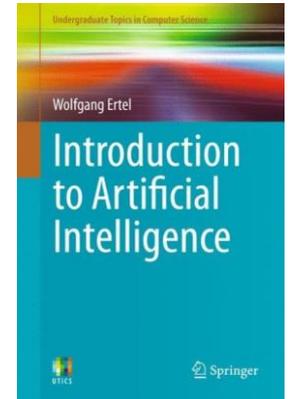
doc2
the 2011
International Joint
Conference on
**Artificial
Intelligence** will
be held in Spain

doc3
apple launches a
new product...



USER PROFILE

| | |
|---------------------|------|
| <u>artificial</u> | 0.11 |
| <u>intelligence</u> | 0.12 |
| apple | 0.20 |
| AI | 0.18 |
| ... | |



Book recommendation

multi-word concepts

Lack of Semantics in User Models

...is not only about polysemy

doc1
AI is a branch of
computer science

doc2
the 2011
International Joint
Conference on
**Artificial
Intelligence** will
be held in Spain

doc3
apple launches a
new product...

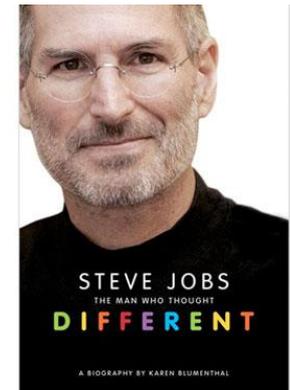
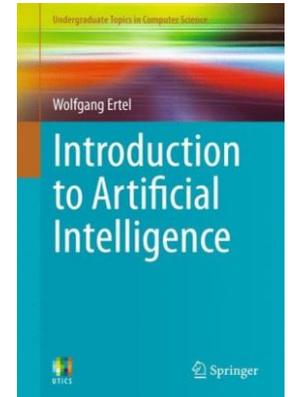


USER PROFILE

| | |
|---------------------|------|
| <u>artificial</u> | 0.11 |
| <u>intelligence</u> | 0.12 |
| apple | 0.20 |
| <u>AI</u> | 0.18 |
| ... | |

Most of the preferences regard AI,
but «apple» is the most relevant
feature in the profile due to
synonymy

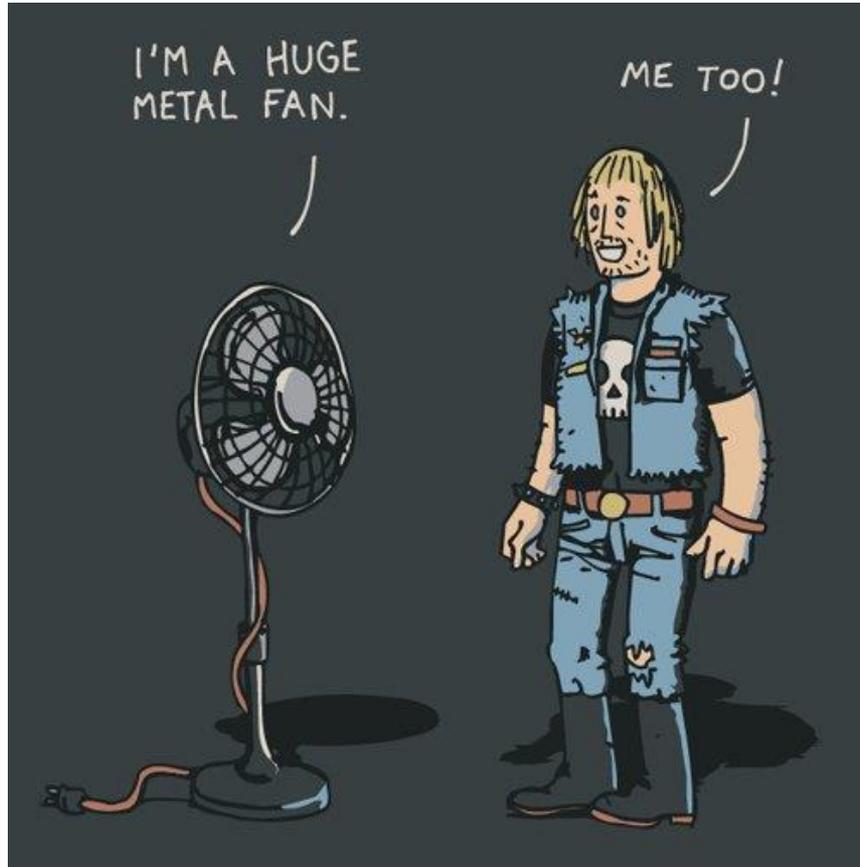
synonymy



Book recommendation

Recap #2

Why do we need semantics?



Because language is inherently ambiguous

- **In general:** to improve **content representation** in intelligent information access platforms
- To avoid typical **issues of natural language representations** (*polysemy, multi-word concepts, synonymy, etc.*)
- To **model user preferences in an effective way**
- To **better understand** the information spread on social media

Agenda

Why?

Why do we need **content**?
Why do we need **semantics**?

How?

How to **introduce semantics**?

Basics of **Natural Language Processing**

Encoding **exogenous semantics**, i.e. *explicit* semantics

Encoding **endogenous semantics**, i.e. *implicit* semantics

What?

Explanation of Recommendations

Conversational Recommender Systems (hands-on)

How to **introduce semantics**?

How to discover **concepts** that connect us to the **the information we are seeking** (*search* task) or **we want to be exposed to** (*recommendation* and *user modeling* tasks)?



We need **some «intelligent» support**
(**as intelligent information access technologies**)

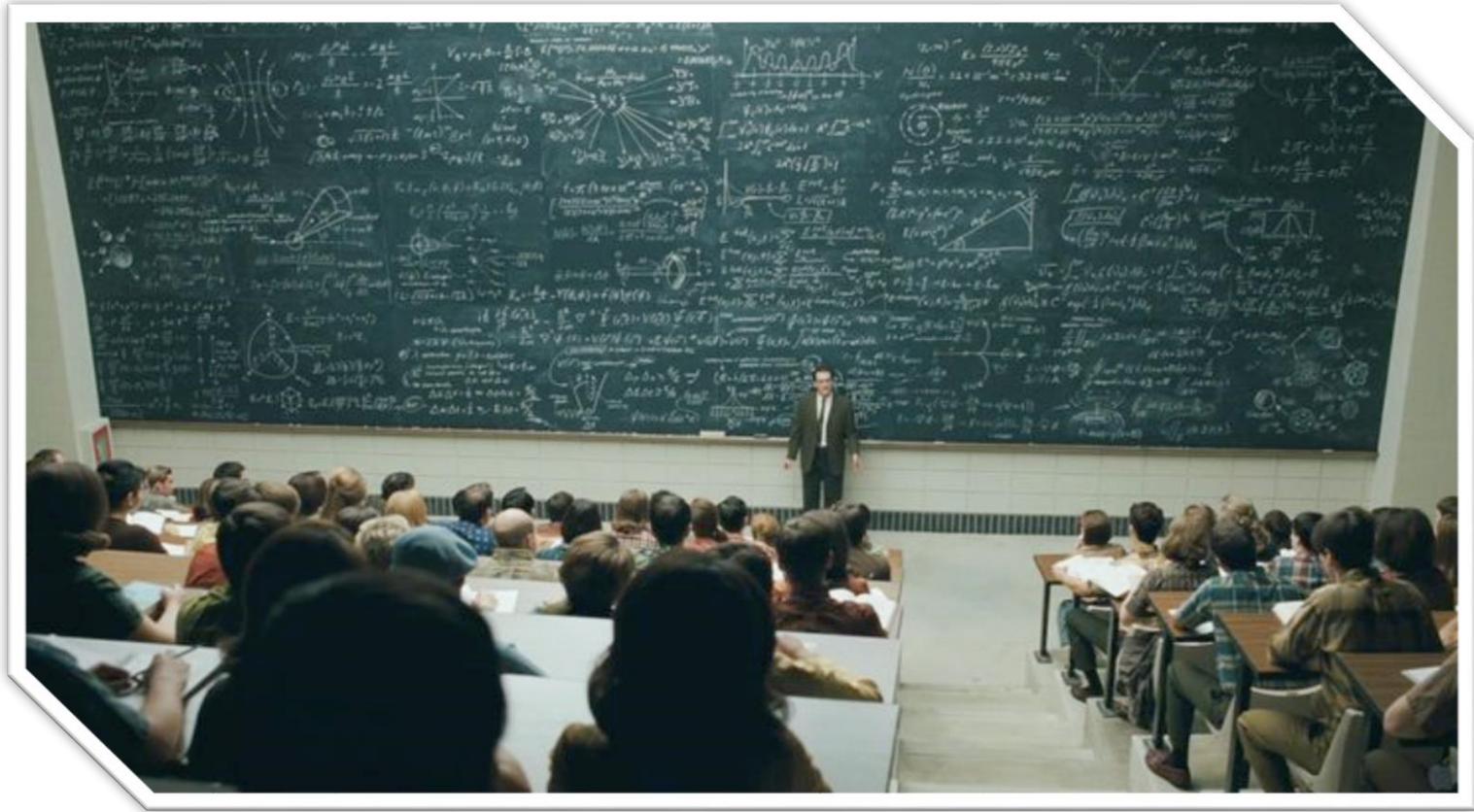


We need to **better understand and represent the content**



...before semantics

some **basics** on



Natural Language Processing (NLP)

Agenda

Why?

Why do we need **content**?
Why do we need **semantics**?

How?

How to **introduce semantics**?

Basics of **Natural Language Processing**

Encoding **exogenous semantics**, i.e. *explicit* semantics

Encoding **endogenous semantics**, i.e. *implicit* semantics

What?

Explanation of Recommendations

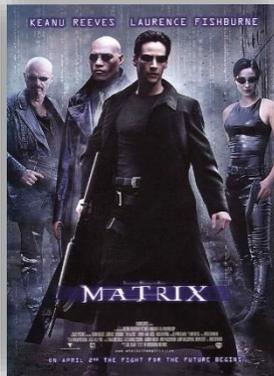
Conversational Recommender Systems (hands-on)

Scenario

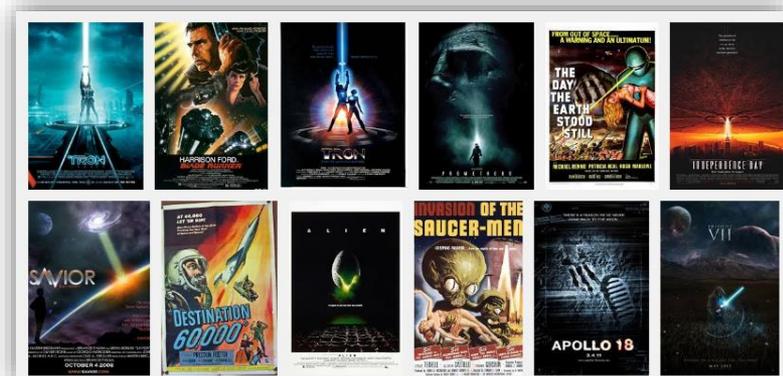
Pasquale really loves the movie «The Matrix», and he asks a content-based recommender system for some suggestions.

Question

How can we **feed the algorithm with some textual features** related to the **movie** to build a **(content-based) profile** and provide recommendations?



Recommendation
Engine



Scenario

The Matrix

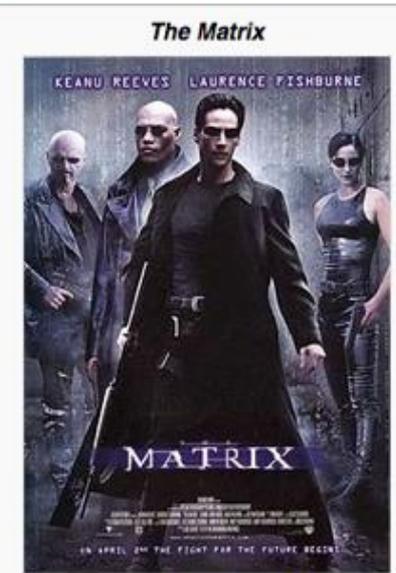
From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see [The Matrix \(franchise\)](#). For other uses, see [Matrix \(disambiguation\)](#).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.^[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's *Allegory of the Cave*,^[6] Jean Baudrillard's *Simulacra and Simulation*^[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.^[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation^[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over \$460 million worldwide. It was generally well-received by critics,^{[10][11]} and won four Academy Awards as well as other accolades including BAFTA



Theatrical release poster

the plot can be a **rich source** of **content-based features**

Scenario

The Matrix

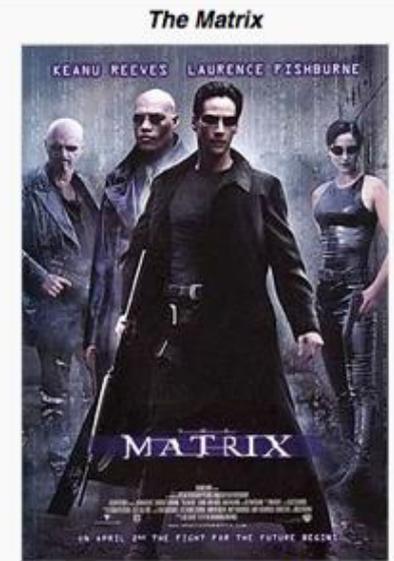
From Wikipedia, the free encyclopedia

This article is about the 1999 film. For the franchise it initiated, see [The Matrix \(franchise\)](#). For other uses, see [Matrix \(disambiguation\)](#).

The Matrix is a 1999 American science fiction action film written and directed by The Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

The Matrix is known for popularizing a visual effect known as "bullet time", in which the heightened perception of certain characters is represented by allowing the action within a shot to progress in slow-motion while the camera's viewpoint appears to move through the scene at normal speed. The film is an example of the cyberpunk science fiction genre.^[5] It contains numerous references to philosophical and religious ideas, and prominently pays homage to works such as Plato's *Allegory of the Cave*,^[6] Jean Baudrillard's *Simulacra and Simulation*^[7] and Lewis Carroll's *Alice's Adventures in Wonderland*.^[8] The Wachowskis' approach to action scenes drew upon their admiration for Japanese animation^[9] and martial arts films, and the film's use of fight choreographers and wire fu techniques from Hong Kong action cinema was influential upon subsequent Hollywood action film productions.

The Matrix was first released in the United States on March 31, 1999, and grossed over \$460 million worldwide. It was generally well-received by critics,^{[10][11]} and won four Academy Awards as well as other accolades including BAFTA



Theatrical release poster

the plot can be a **rich source** of **content-based features**

...but we need to **properly process it** through a pipeline of **Natural Language Processing (NLP)** techniques

Basic NLP operations for lexical analysis

- ✓ **Normalization** strips unwanted characters/markup (e.g. HTML/XML tags, punctuation, numbers, etc.)
- ✓ **Tokenization** breaks text into tokens
- ✓ **Stop words removal** excludes common words having little semantic content
- ✓ **Lemmatization** reduces inflectional/variant forms to base form (lemma in the dictionary)
am, are, is, was, were → **be**
- ✓ **Stemming** reduce terms to their “roots”
automate(s), automatic, automation → **automat**

vocabulary

Example

The Matrix is a 1999 American-Australian neo-noir science fiction action film written and directed by the Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called "the Matrix", created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer "Neo" learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the "dream world".

Normalization

strips unwanted characters/markup (e.g. HTML/XML tags, punctuation marks, quotation marks, numbers, hyphens, etc.)

The Matrix is a 1999 American-Australian neo-noir science fiction action film written and directed by the Wachowskis, starring Keanu Reeves, Laurence Fishburne, Carrie-Anne Moss, Hugo Weaving, and Joe Pantoliano. It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called the Matrix, created by sentient machines to subdue the human population, while their bodies' heat and electrical activity are used as an energy source. Computer programmer Neo learns this truth and is drawn into a rebellion against the machines, which involves other people who have been freed from the dream world.

Tokenization

breaks text into tokens

The Matrix is a 1999 American Australian neo noir science fiction action film written and directed by the Wachowskis starring Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving and Joe Pantoliano It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called the Matrix created by sentient machines to subdue the human population while their bodies heat and electrical activity are used as an energy source Computer programmer Neo learns this truth and is drawn into a rebellion against the machines which involves other people who have been freed from the dream world

Tokenization issues

compound words

- science-fiction: break up **hyphenated** sequence?
- Keanu Reeves: **one token or two**? How do you decide it is one token?

numbers and dates

- (800) 234-2333
- 3/20/91 Mar. 20, 1991 20/3/91
- 55 B.C.

Tokenization issues

language issues

- German compound nouns are not segmented
Lebensversicherungsgesellschaftsangestellter means
life insurance company employee
- Chinese, Japanese and Korean have no spaces (not always guaranteed a unique tokenization)

莎拉波娃现在居住在美国东南部的佛罗里达

- Arabic or Hebrew is basically written right to left, but with certain items (like numbers) written left to right

استقلت الجزائر في سنة 1962 بعد 132 عام من الاحتلال الفرنسي.

Algeria achieved its independence in 1962 after 132 years of French occupation

Stop words removal

excludes common words having little semantic content

The Matrix is a 1999 American Australian neo noir science fiction action film written and directed by the Wachowskis starring Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving and Joe Pantoliano It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called the Matrix created by sentient machines to subdue the human population while their bodies heat and electrical activity are used as an energy source Computer programmer Neo learns this truth and is drawn into a rebellion against the machines which involves other people who have been freed from the dream world

Stop words removal

The Matrix is a 1999 American Australian neo noir science fiction action film written and directed by the Wachowskis starring Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving and Joe Pantoliano It depicts a dystopian future in which reality as perceived by most humans is actually a simulated reality called the Matrix created by sentient machines to subdue the human population while their bodies heat and electrical activity are used as an energy source Computer programmer Neo learns this truth and is drawn into a rebellion against the machines which involves other people who have been freed from the dream world

Lemmatization

reduces inflectional/derivational/variant forms
to base form (lemma in the dictionary)

am, are, is, was, were → **be**

The Matrix is a 1999 American Australian neo noir science fiction action film **written** and **directed** by the Wachowskis **starring** Keanu Reeves Laurence Fishburne Carrie Anne Moss Hugo Weaving and Joe Pantoliano It **depicts** a dystopian future in which reality as **perceived** by most **humans** is actually a **simulated** reality **called** the Matrix **created** by sentient **machines** to subdue the human population while their **bodyies** heat and electrical activity are **used** as an energy source Computer programmer Neo **learns** this truth and is **drawn** into a rebellion against the **machines** which **involves** other people who have been **freed** from the dream world

Example

Matrix 1999 American Australian neo noir science fiction
action film write direct Wachowskis star Keanu Reeves
Laurence Fishburne Carrie Anne Moss Hugo Weaving
Joe Pantoliano depict dystopian future reality perceived
human simulate reality call Matrix create sentient
machine subdue human population body heat electrical
activity use energy source Computer programmer Neo
learn truth draw rebellion against machine involve people
free dream world

**next step: assign a weight to each feature
(e.g., through TF-IDF)
to identify the most relevant ones**

Weighting features (= terms): TF-IDF

terms frequency – inverse document frequency

best known weighting scheme in information retrieval.

Weight of a term as product of **tf weight** and **idf weight**

$$w_{t,d} = (1 + \log \text{tf}_{t,d}) \times \log(N / \text{df}_t)$$

tf **number of times** the term occurs in the document

idf depends on **rarity** of a term in a collection

tf-idf increases with the number of occurrences within a document, as well as with the rarity of the term in the collection

Example

Matrix 1999 American Australian neo noir science fiction
action **film** write direct Wachowskis star Keanu Reeves
Laurence Fishburne Carrie Anne Moss Hugo Weaving
Joe Pantoliano depict **dystopian** future reality
perceived human simulate reality call Matrix create
sentient machine subdue human population body heat
electrical activity **use** energy source Computer
programmer Neo learn truth draw **rebellion** against
machine involve people free dream world

green=high IDF

red=low IDF

The Matrix representation

Matrix

1999

American

Australian

science

fiction

Hugo

...

world

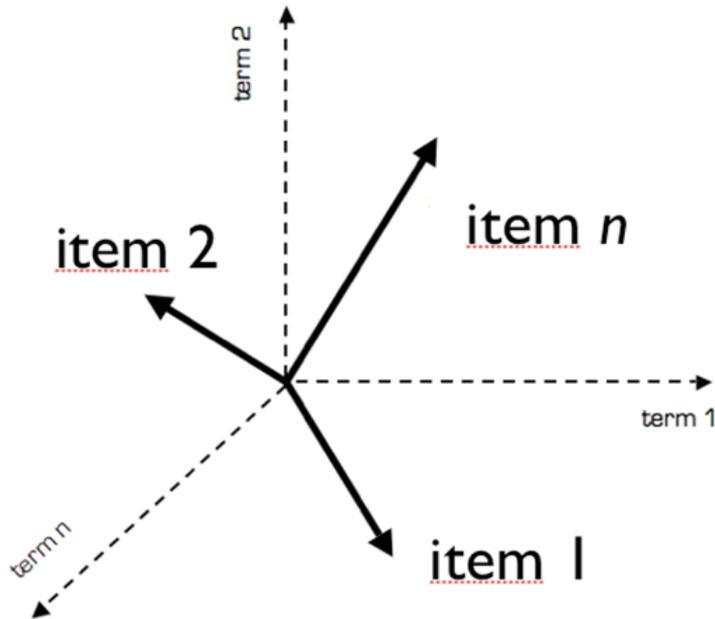
keywords

a portion of Pasquale's
content-based profile



given a content-based profile, we
can **easily build a basic
recommender system** through
Vector Space Model and
similarity measures

Vector Space Model (VSM)



given a set of n features
(terms in the vocabulary)

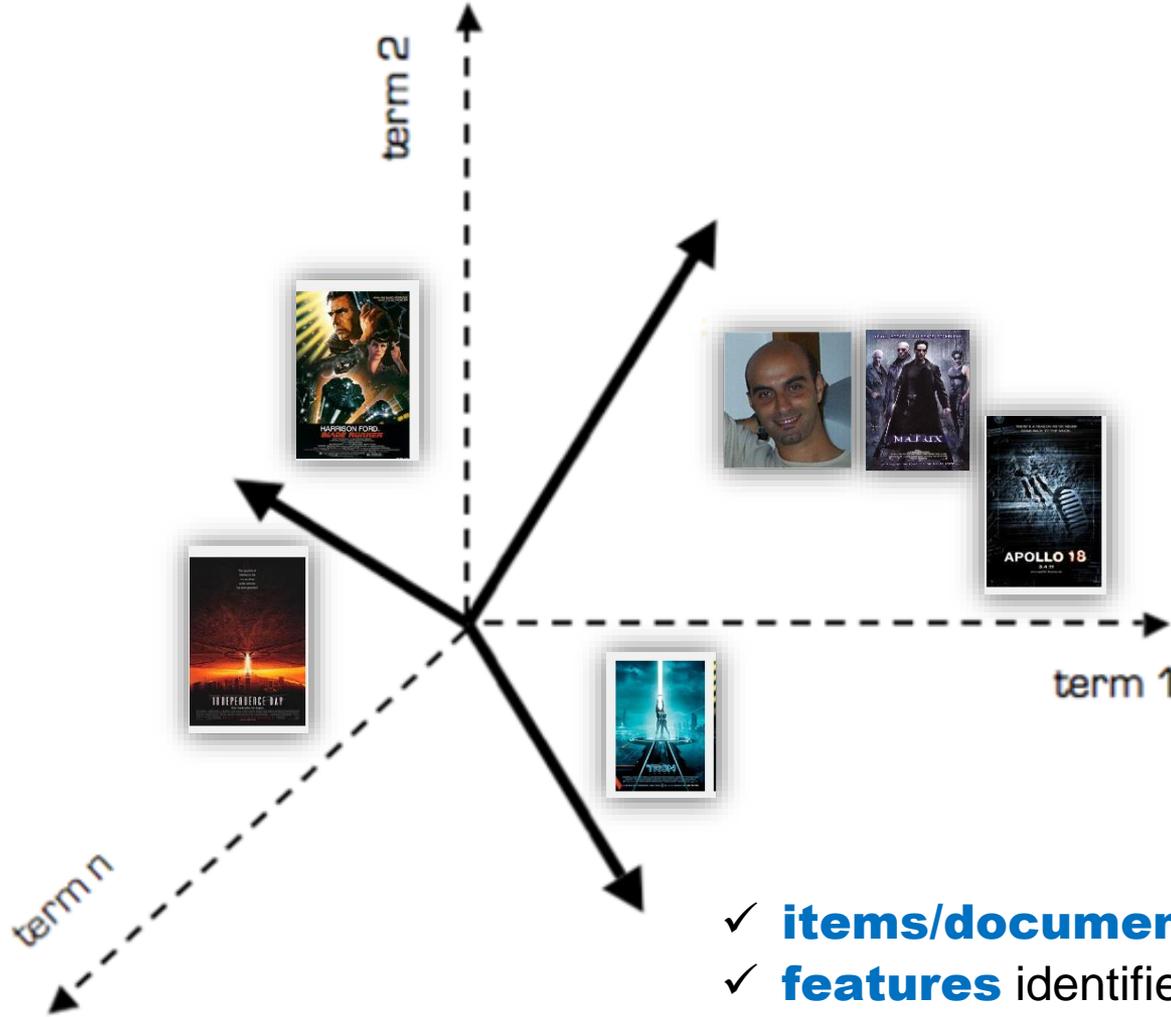
$$f = \{f_1, f_2, \dots, f_n\}$$

each item i represented as a point in a
 n -dimensional vector space

$$I = (w_{f1}, w_{f2}, \dots, w_{fn})$$

w_{fi} is the **weight** of feature i
in the item I

Basic Content-based Recommendations

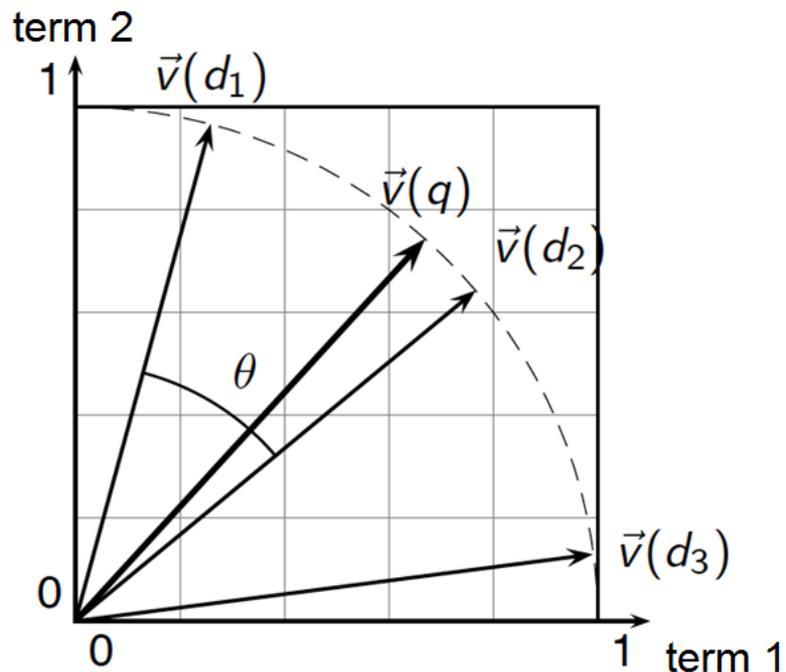


- ✓ **items/documents** represented as **vectors**
- ✓ **features** identified through **NLP operations**
- ✓ **features** weighted using **tf-idf**
- ✓ **cosine measure** for computing similarity between vectors

Similarity between vectors

cosine similarity

$$\cos(\vec{I}, \vec{J}) = \frac{\overset{\text{dot product}}{\vec{I} \bullet \vec{J}}}{|\vec{I}| |\vec{J}|} = \frac{\overset{\text{unit vectors}}{\frac{\vec{I}}{|\vec{I}|}} \bullet \frac{\vec{J}}{|\vec{J}|}}{1 \cdot 1} = \frac{\sum_{i=1}^{|\mathcal{V}|} I_i J_i}{\sqrt{\sum_{i=1}^{|\mathcal{V}|} I_i^2} \sqrt{\sum_{i=1}^{|\mathcal{V}|} J_i^2}}$$



Basic Content-based Recommendations

Drawbacks

Matrix

1999

American

Australian

science

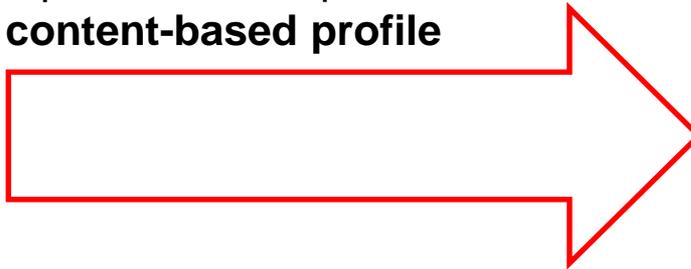
fiction

Hugo

...

world

a portion of Pasquale's
content-based profile



Recommendation:
Notre Dame de Paris,
by Victor Hugo



Why?

Entities as «Hugo Weaving» were not modeled

Basic Content-based Recommendations

Drawbacks

Matrix

1999

American

Australian

science

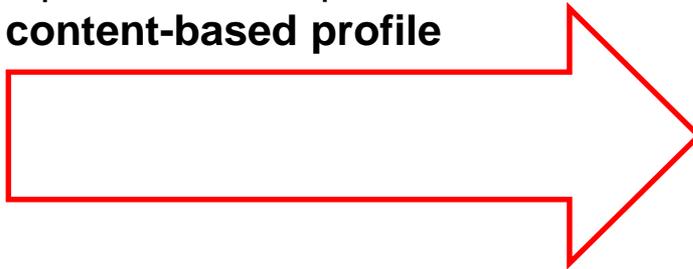
fiction

Hugo

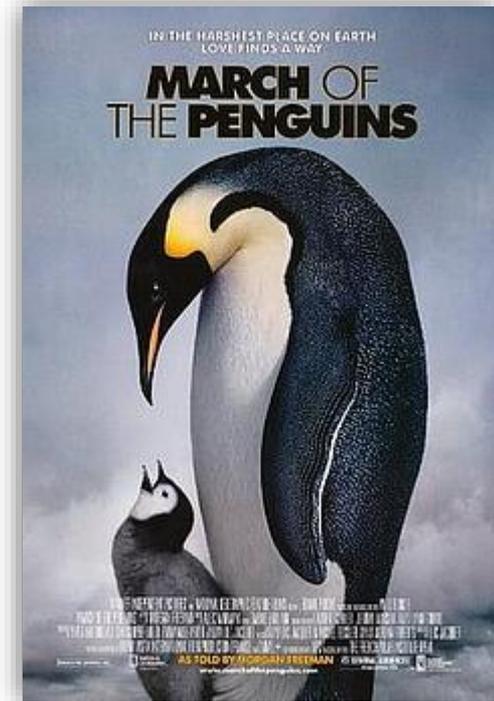
...

world

a portion of Pasquale's
content-based profile



Recommendation:
The March of Penguins



Why?

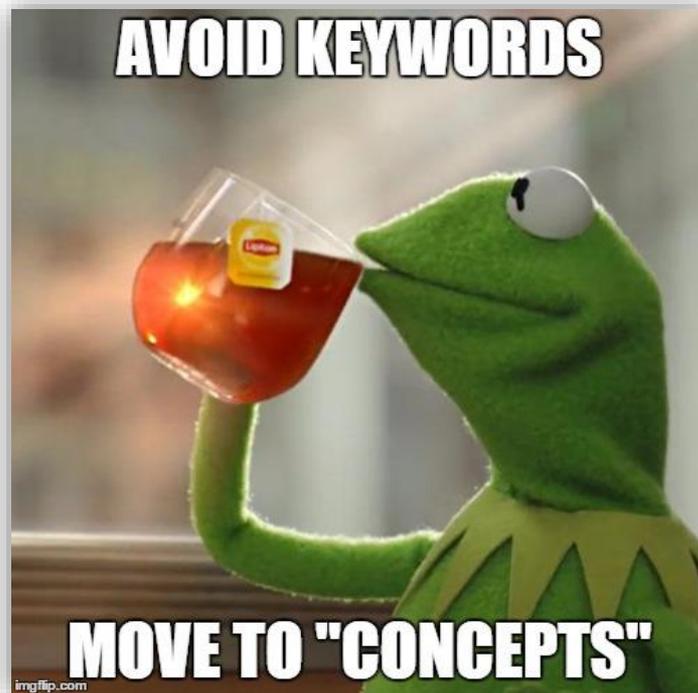
**More complex concepts
as «science fiction»
were not modeled as
single features**

Basic Content-based Recommendations

Vision



Bad recommendations



Recap #3

basics of **NLP** and **keyword**-based representation



- **Natural Language Processing (NLP) techniques necessary** to build a content-based profile
- **basic Content-based Recommender Systems can be easily built** through VSM and TF-IDF
- **keyword-based representation too poor** and can drive to bad modeling of preferences (and bad recommendations)
- **we need to shift from keywords to concepts**