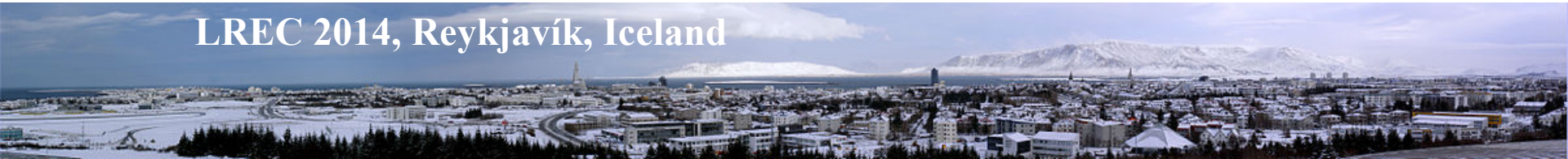
Thematic Cohesion: Measuring Terms Discriminatory Power Toward Themes

Clément de Groc
Xavier Tannier
Claude de Loupy



Thematic lexicons

Domain-specific lists of terms (*e.g.* Astronomy)

observatory
cosmology
amateur
equatorial
planets
hubble
gallery
located
spotting
lunar
ccd
outreach
x-ray
astronomers
comets
sun
galaxies
solar
space
newtonian
supernovae
data
sky
meetings
eclipses
software
refractors
microscopes
universe
astrophysics
dobsonian
physics
nasa
missions
astrophotography
planetarium
satellites
earth
cosmic
mounts
mars
astro
seti
research
asteroid
meteor
telescopes
meteorites
extrasolar
neutron
objects
constellations
pulsars
nebulae
astronomy
pluto
scopes
binoculars
image
accessories
optical
eyepieces
celestial
celestron
cameras
clusters
equipment
science
night
observation
star
manufactures
auroral
stellar
system
moon
planetary
deep
kuiper
focusers
meade

Thematic lexicons

Domain-specific lists of terms (*e.g.* Astronomy)

observatory
cosmology
amateur
equatorial
planets
hubble
gallery
located
spotting
lunar
ccd
outreach
x-ray
astronomers
comets
sun
galaxies
solar
space
newtonian
supernovae
data
sky
meetings
eclipses
software
universe
refractors
microscopes
astrophysics
dobsonian
physics
nasa
missions
astrophotography
planetarium
satellites
earth
cosmic
mounts
mars
astro
seti
research
asteroid
meteor
telescopes
meteorites
extrasolar
neutron
objects
constellations
pulsars
nebulae
astronomy
pluto
scopes
binoculars
image
accessories
optical
eyepieces
celestial
celestron
cameras
clusters
equipment
science
night
observation
star
manufactures
auroral
stellar
system
planetary
deep
moon
kuiper
focusers
meade

Thematic lexicons

Domain-specific lists of terms (*e.g.* Astronomy)

observatory
amateur
equ
planets
hubble
lunar
cc
x-ray
astronomers
solar
cor
data
meetings
eclipses
microscopes
physics
astrophotography
earth
cosmic
m
research
meteorites
objects
astronomy
pluto
opti
celestron
science
ob
manufactures
a
system
planetary
focusers
meade

- Building a domain-specific, thematic lexicon is **long** and **expensive**
- Semi-automatic processes exists (Baroni and Bernardini, 2004; Kilgarriff and Grefenstette, 2003; Wang and Cohen, 2007)
- However, a **manual validation** step is still required when high quality resources are needed

Objectives

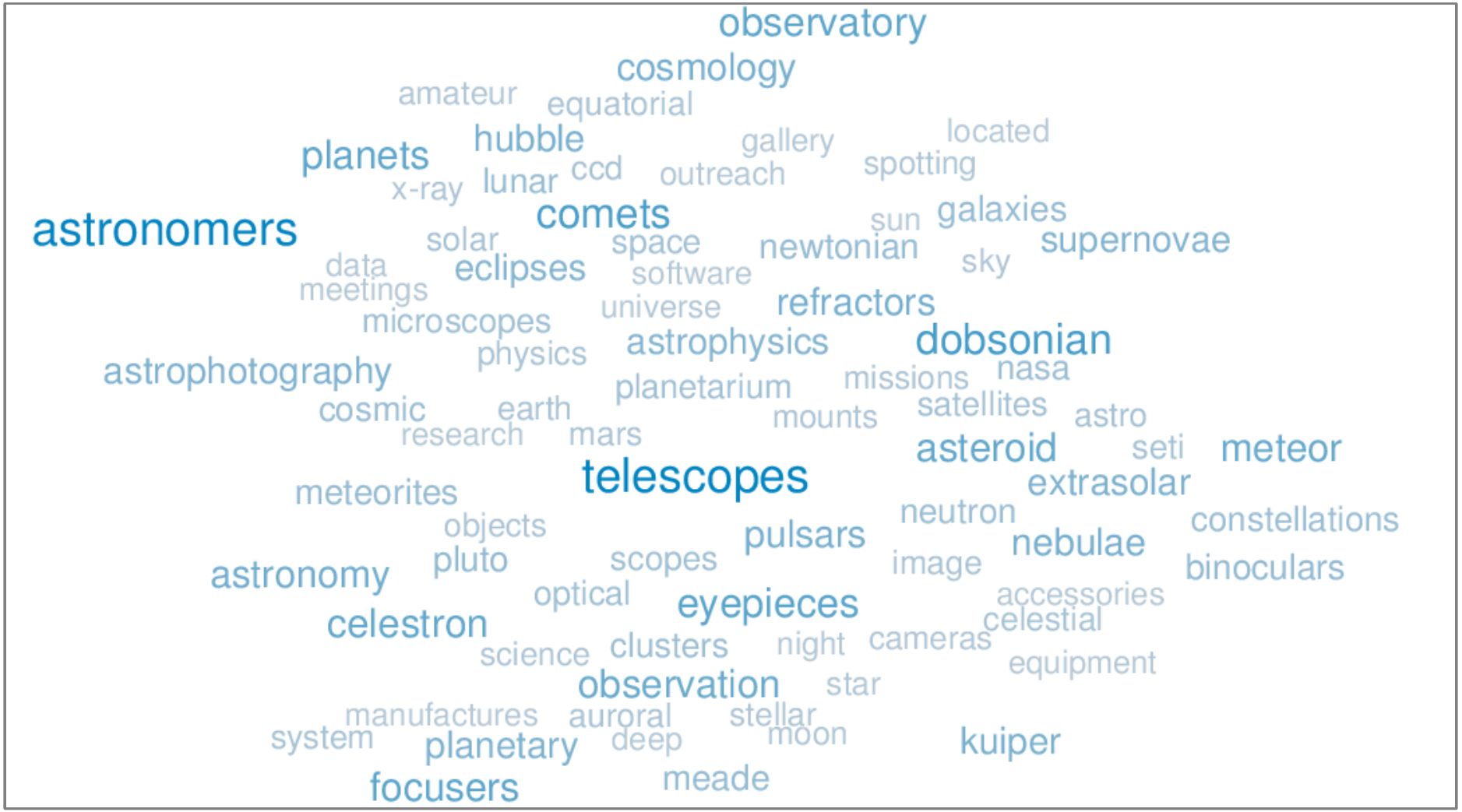
Objectives

Input: a thematic lexicon

observatory
cosmology
amateur
equatorial
planets
hubble
gallery
located
spotting
lunar
ccd
outreach
x-ray
astronomers
comets
sun
galaxies
solar
space
newtonian
supernovae
data
sky
meetings
eclipses
software
refractors
microscopes
universe
astrophysics
dobsonian
physics
nasa
missions
astrophotography
planetarium
satellites
earth
cosmic
mounts
astro
seti
mars
research
asteroid
meteor
telescopes
meteorites
extrasolar
neutron
objects
constellations
pulsars
nebulae
astronomy
pluto
scopes
binoculars
image
accessories
optical
eyepieces
celestial
celestron
cameras
clusters
equipment
science
night
observation
star
manufactures
auroral
stellar
system
moon
planetary
deep
kuiper
focusers
meade

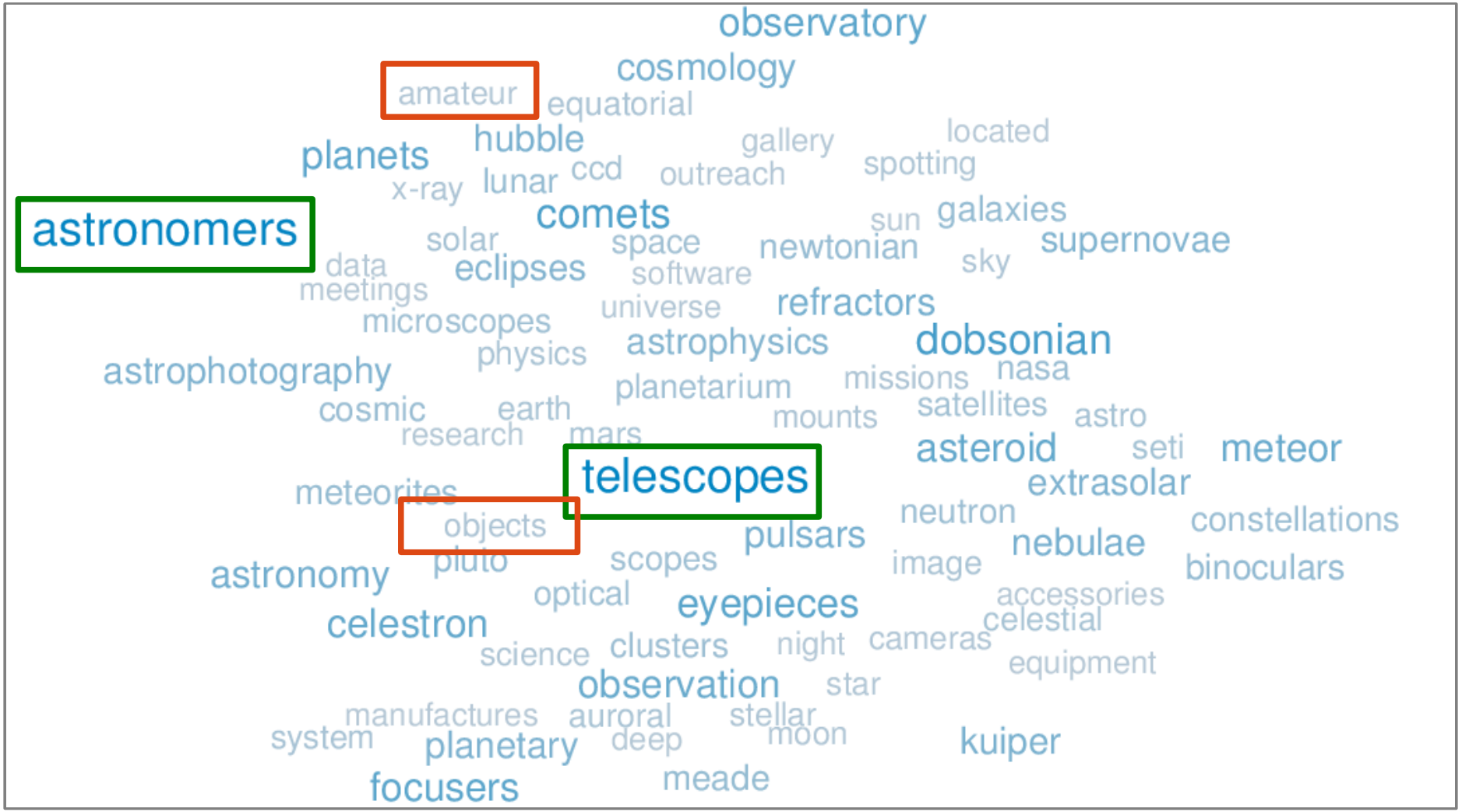
Objectives

Output: a weighted thematic lexicon



Objectives

Output: a weighted thematic lexicon



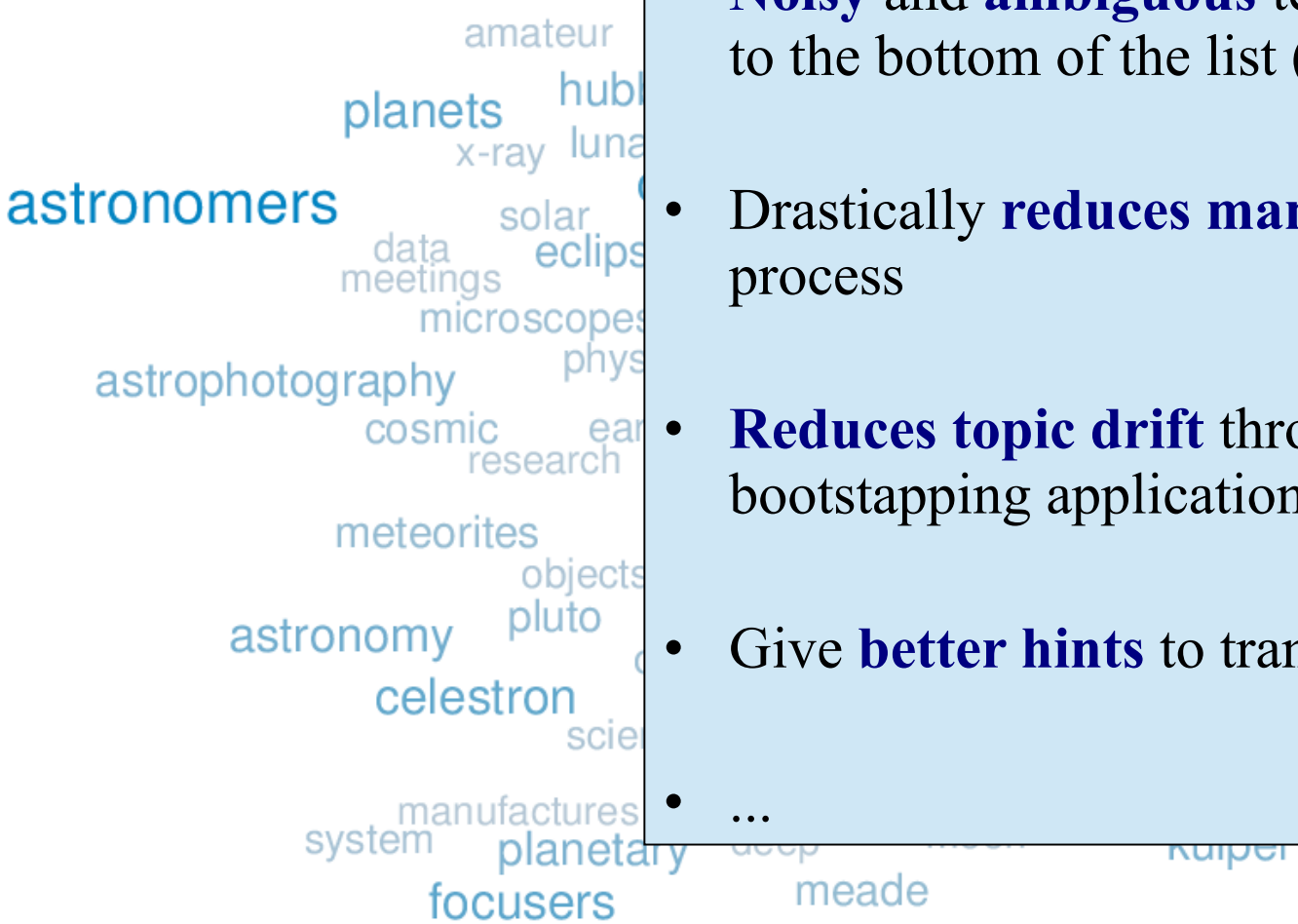
Objectives

Input: a thematic lexicon

Output: term weights

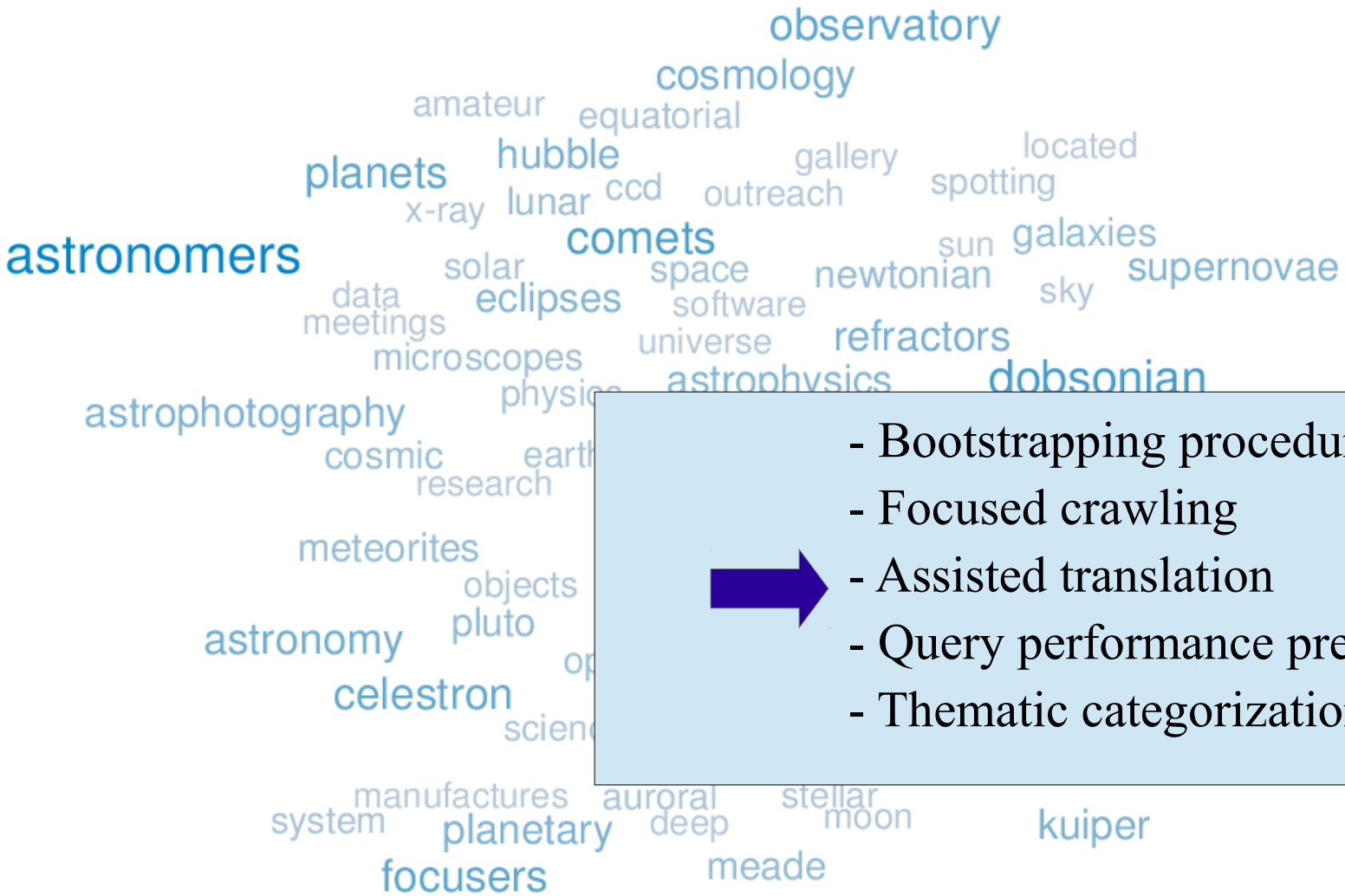
$$L_T = (t_1, t_2, \ldots, t_N) \Rightarrow w_{L_T} = (w_1, w_2, \ldots, w_N)$$

Objectives



- **Noisy** and **ambiguous** terms are relegated to the bottom of the list (low weights)
- Drastically **reduces manual validation** process
- **Reduces topic drift** through iteration in bootstapping applications
- Give **better hints** to translators
- ...

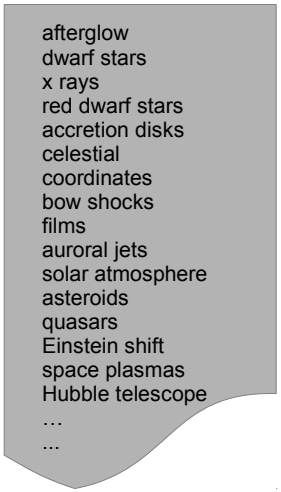
Objectives



- Bootstrapping procedures
- Focused crawling
- Assisted translation
- Query performance prediction
- Thematic categorization

Thematic Cohesion Value

Collecting Data



Term t_i → Standard (general) web search engine (blekko) → M top results → Corpus C_i

$\ldots t_1 \ldots t_4$
$\ldots\ldots t_5 \ldots$
$\ldots t_i \ldots t_i$
$\ldots\ldots t_8 \ldots\ldots$

either the **entire web pages**
or the **search engine snippets**

Graph Representation

Corpus C_i

$\ldots t_1 \ldots t_4 \ldots$

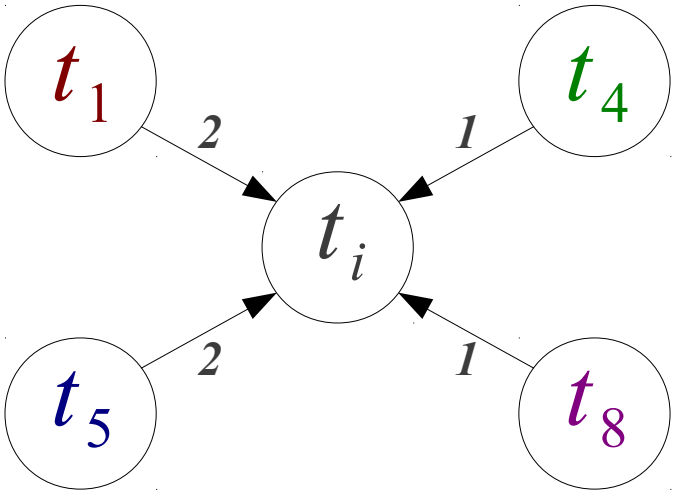
$\ldots\ldots t_5 \ldots\ldots\ldots$

$\ldots t_i \ldots\ldots\ldots t_i$

$\ldots\ldots t_8 \ldots\ldots\ldots$

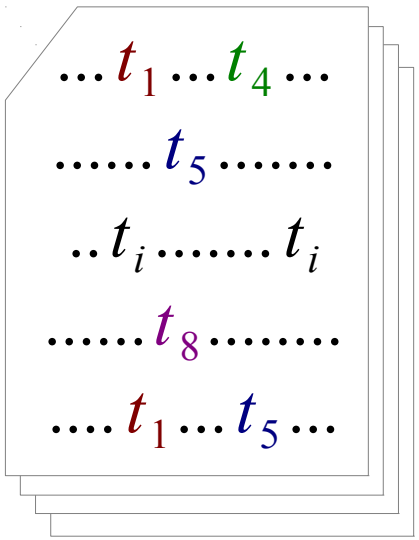
$\ldots\ldots t_1 \ldots t_5 \ldots$

Cooccurrence Graph

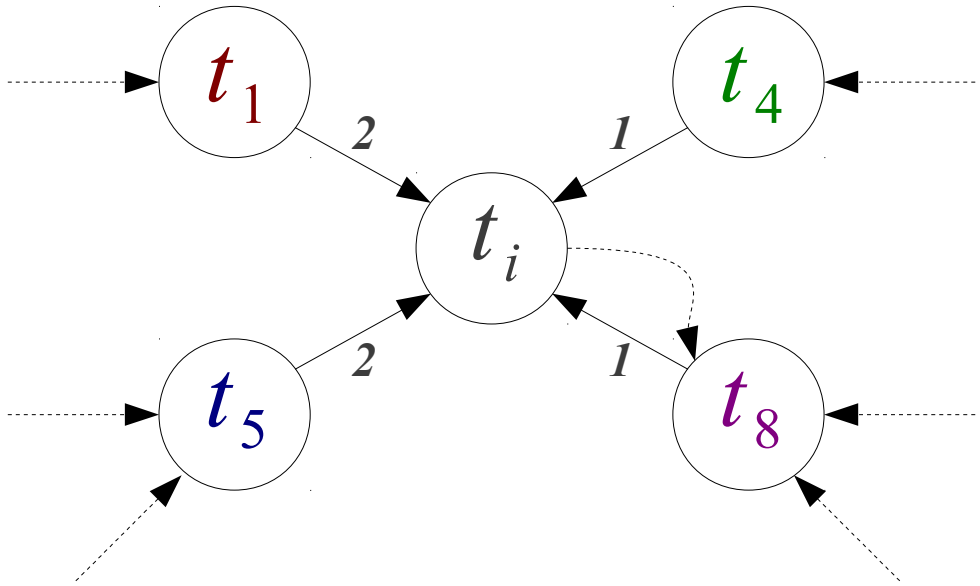


Graph Representation

Corpora $C_{1..n}$



Cooccurrence Graph



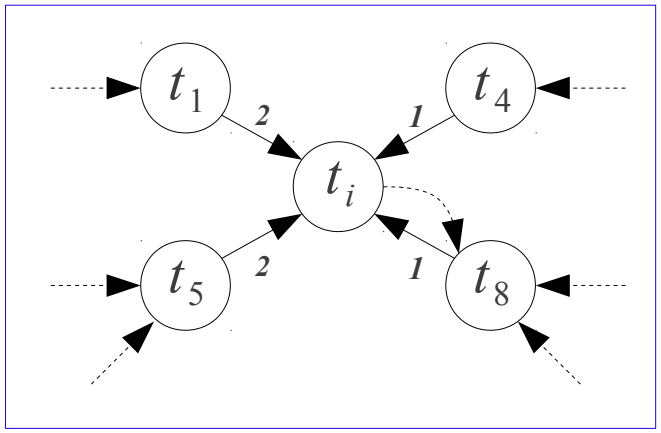
Thematic Cohesion Value

- **Thematic Cohesion Value** of a word =
  How ***central*** this word is in the lexicon cooccurrence graph

- Definition #1: In-degree

Score of term t_i = Number of occurrences of lexicon terms in C_i

$$w_i = \sum_{t_j} n_{t_j, C_i}$$

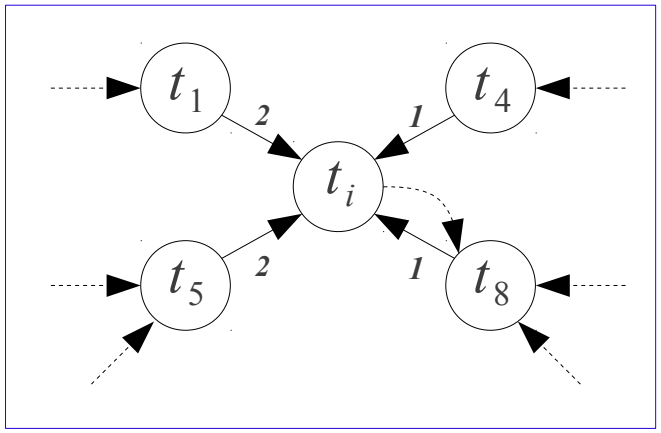


Thematic Cohesion Value

- **Thematic Cohesion Value** of a word =
  How ***central*** this word is in the lexicon cooccurrence graph

- Definition #2: Normalized in-degree (score in [0,1])

$$w_i = \frac{\sum_{t_j} n_{t_j, C_i}}{\sum_{k \in 1..n} \sum_{t_j} n_{t_j, C_k}}$$

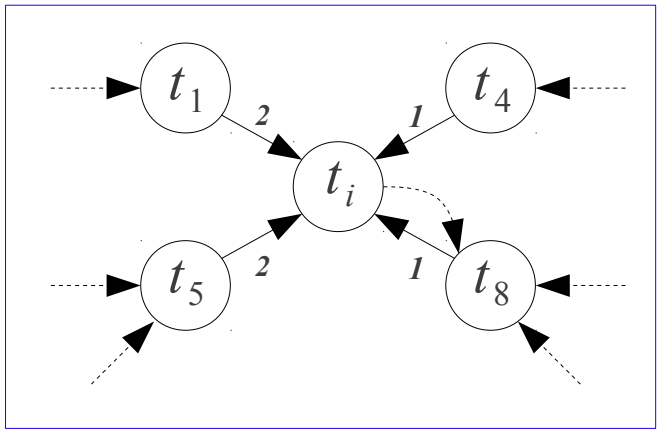


Thematic Cohesion Value

- **Thematic Cohesion Value** of a word =
  How ***central*** this word is in the lexicon cooccurrence graph

- Definition #3: Consider other terms weights (**random walk**)

$$w_i = \sum_{t_j} \frac{n_{t_j, C_i} \times \boldsymbol{w_j}}{\sum_{k \in 1..n} n_{t_j}, C_k}$$

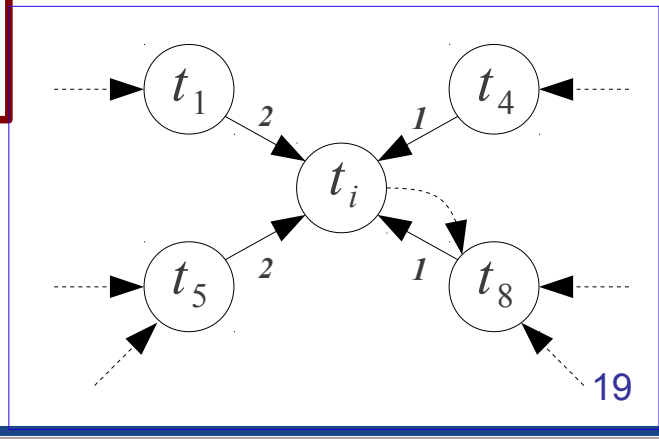


Thematic Cohesion Value

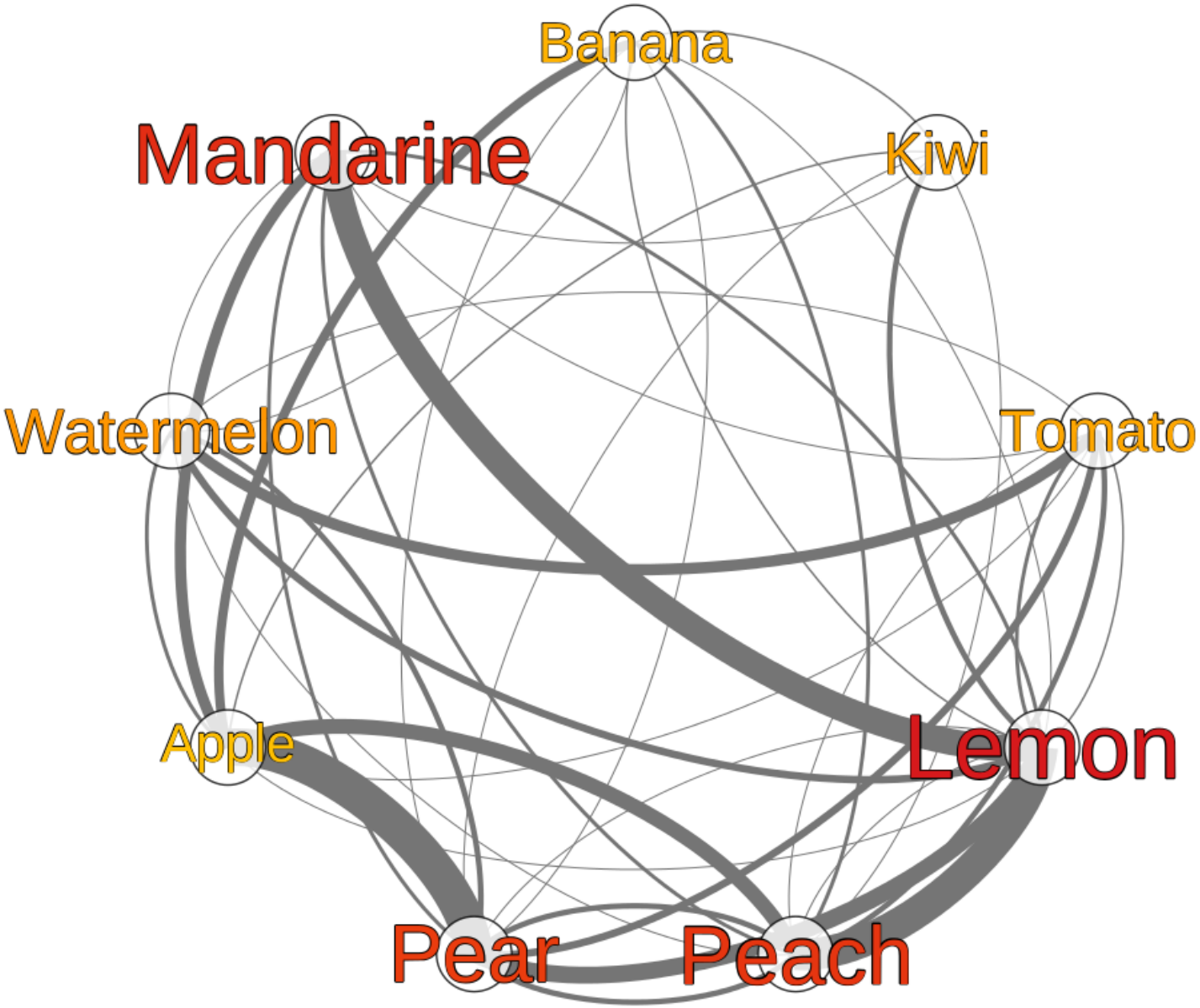
- **Thematic Cohesion Value** of a word =
  How ***central*** this word is in the lexicon cooccurrence graph

- Final Definition: Convergence condition → add teleportation vector (**PageRank**)

$$w_i = \frac{(1-\alpha)}{N} + \alpha . \sum_{t_j} \frac{n_{t_j, C_i} \times w_j}{\sum_{k \in 1..n} n_{t_j, C_k}}$$



Thematic Cohesion Value



Evaluation

Evaluation

- **Evaluation** of:
  - **Behavior**
    - Influence of the number of documents M
    - Web pages vs. snippets
    - Influence of the initial lexicon size N
  - **Relevance**
    - Best terms should lead to more precise documents for the topic

Evaluation : Measure Behavior

Reference lexicons:

- ***Astronomy*** (2940 terms, the Astronomy Thesaurus)
- ***Statistics*** (2752 terms, the ISI Glossary)
- ***Medical*-1** (2000 terms, MeSH)
- ***Medical*-2** (2000 terms, MeSH)

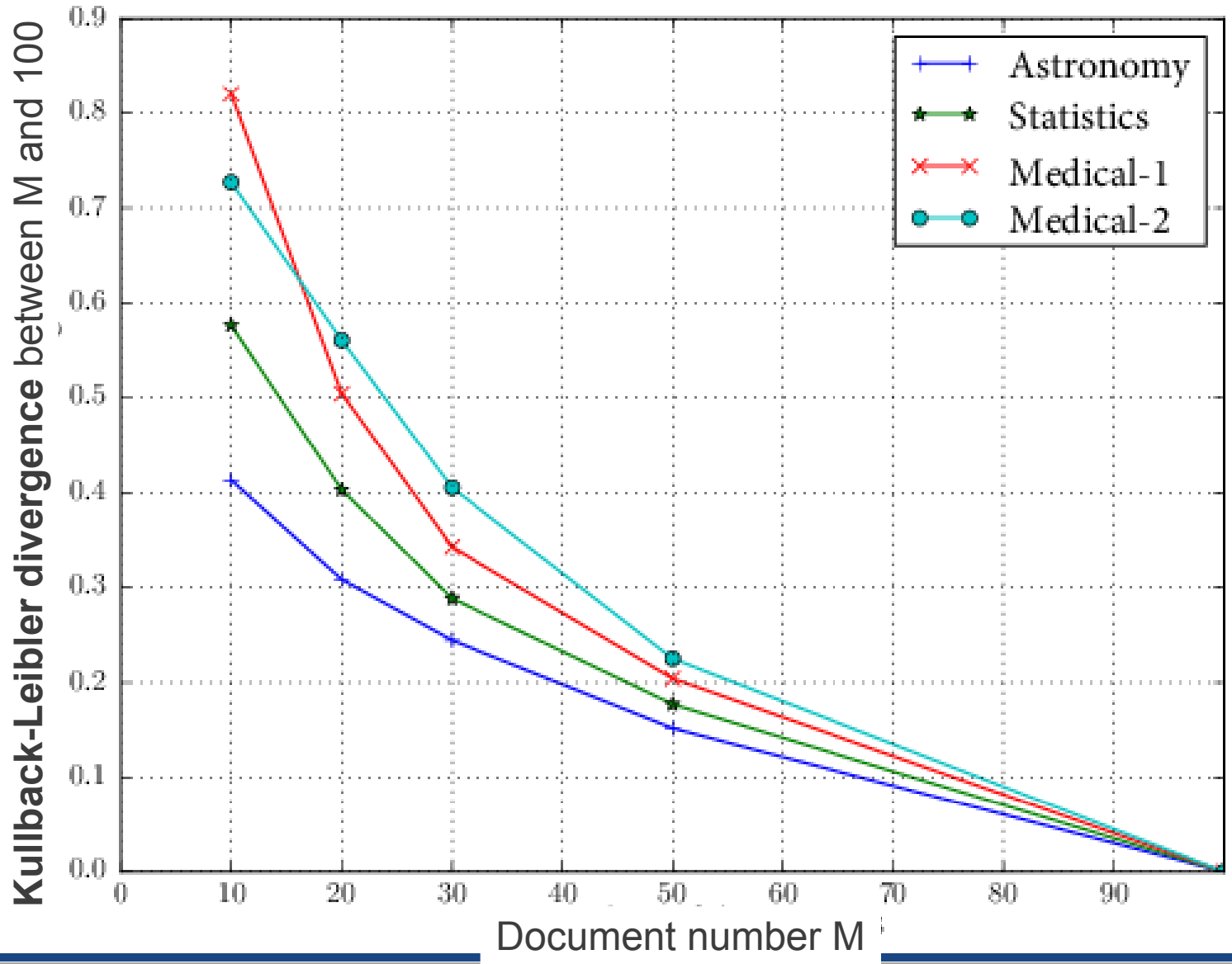
Evaluation : Measure Behavior

- What is the **lowest sufficient number of documents** (to obtain stable results) ?

- Study when **results stop changing by adding documents**
  - Different values of M = number of documents
  - max(M) = 100 web pages or 500 snippets
  - KL-Divergence between lists obtained with M and max(M)

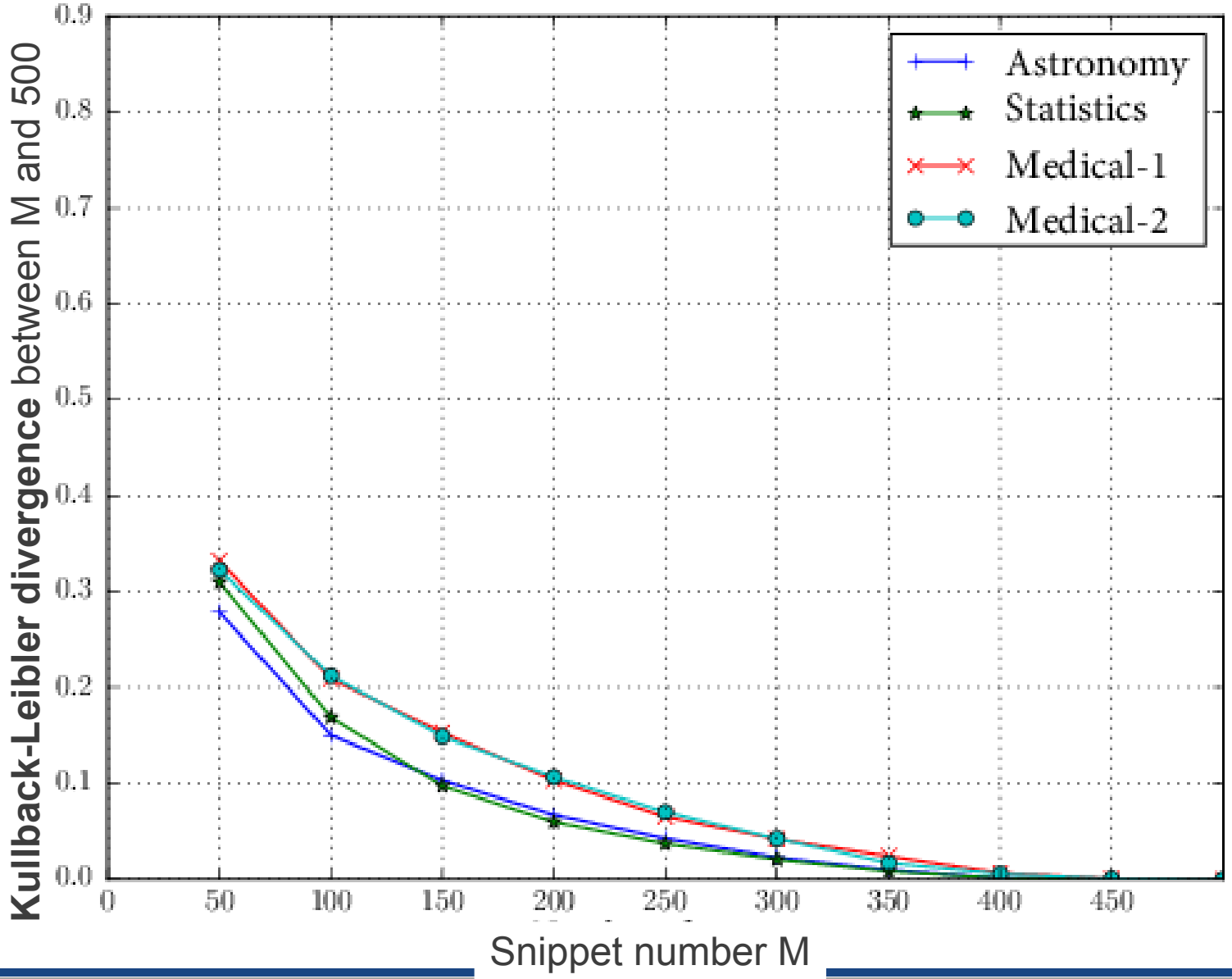
Evaluation : Measure Behavior

Web Pages



Evaluation : Measure Behavior

Snippets



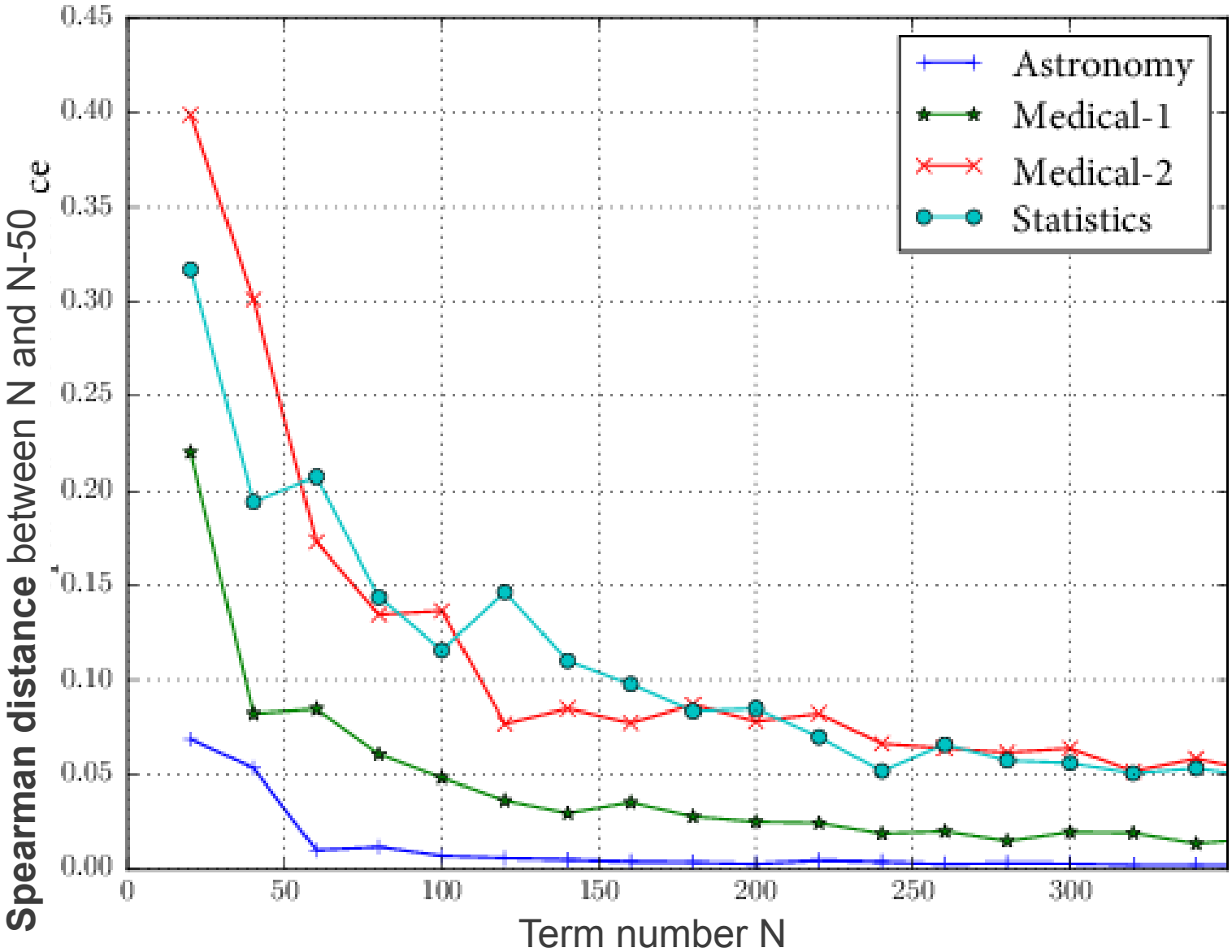
Evaluation : Measure Behavior

- What is the **lowest sufficient number of terms in the lexicon** (to obtain stable results) ?

- Study when **ranking stops being messed up by adding new terms** to the lexicon
  - Different values of N = number of terms (between 20 and 1000)
  - Spearman distance of the ranked lists with N and N+50 terms.

Evaluation : Measure Behavior

Terms



Evaluation : Relevance

- OpenDirectory (DMOZ)

Categories

dmoz — In partnership with Aol.

Follow @dmoz | about dmoz | dmoz blog | suggest URL | help | link | editor login

Search *advanced*

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

Science
Biology, Psychology, Physics...

Shopping
Clothing, Food, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

World
Català, Česky, Dansk, Deutsch, Español, Esperanto, Français, Galego, Hrvatski, Italiano, Lietuvių, Magyar, Nederlands, Norsk, Polski, Português, Română, Slovensky, Suomi, Svenska, Türkçe, Български, Ελληνικά, Русский, Українська, العربية, עברית, ไทย, 日本語, 简体中文, 繁體中文, ...

Become an Editor | Help build the largest human-edited directory of the web

4,220,543 sites - 89,489 editors - over 1,021,395 categories

Evaluation : Relevance

- OpenDirectory (DMOZ)

dmoz

Follow @dmoz

about dmoz | dmoz blog | suggest URL | update listing

Search | the entire directory

Top: Science: Social Sciences: Linguistics: Computational Linguistics: **Conferences** *(33)*

- 2000 *(3)*
- 2001 *(5)*
- 2002 *(3)*
- 2003 *(6)*
- 2004 *(3)*
- 2005 *(0)*
- 2006 *(0)*
- 2007 *(0)*
- 2008 *(0)*
- pre-2000 *(5)*

See also:

- Computers: Artificial Intelligence: Conferences and Events *(52)*
- Computers: Computer Science: Conferences *(179)*
- Computers: Human-Computer Interaction: Conferences *(38)*
- Science: Social Sciences: Linguistics: Conferences *(91)*

- 5th international Conference of the Global WordNet Association - The GWC-2010 will be held at IIT Bombay, Mumbai, India, during 31st Jan - 4th Feb, 2010.
- The Association for Computational Linguistics: Conferences - Information on upcoming ACL (and associated) conferences. Also archives and mirrors of past conferences.
- Conferences and Workshops - Listed by year, and maintained by the community.
- LREC Conferences - The International Conference on Language Resources and Evaluation is organised by ELRA biennially with the support of institutions and organisations involved in HLT. large number of people working and interested in HLT.

Site description

Evaluation : Relevance

- **Resources**:
  - Index sites from 340 categories of DMOZ (second-level)
  - For each category, build a lexicon with 200 best tf.idf terms from all site descriptions
- **Methodology**:
  - Compute Thematic Cohesion Values for each lexicon
  - Issue each term as a query to our OpenDirectory search engine
  - **Idea**: *the more a term is relevant to a topic, the better the precision of retrieved documents*

    → **Compute** the average precision of the set of retrieved documents.

    → **Compare** average precision and thematic cohesion scores (Spearman coefficient)

Evaluation : Relevance

Measure	Spearman ρ	Significance (% of categories where p-value < 0.05)
tf.idf	0.200	32%
Thematic cohesion	0.434	74%

Conclusion

Conclusion

- A novel Thematic Cohesion Measure
- Weights thematic lexicon terms according to their discriminatory power toward the theme

- Use of a general search engine
- Snippets are more robust and as relevant as web pages

- Useful for corpora bootstrapping, assisted translation, query performance prediction, etc.