Tagging Scientific Publications Using Wikipedia and NLP Tools

Comparison on the ArXiv dataset

Michał Łopuszyński , Łukasz Bolikowski

Agenda

- **What? Why? How?**

  Motivation, dataset, details of the two employed tagging methods, first based on Wikipedia (WIKI) and second based on noun phrases (NP)

- **Comparison of the WIKI and NP based method**

  Weaknesses and strengths of both methods by example

- **Statistical properties of obtained tags**

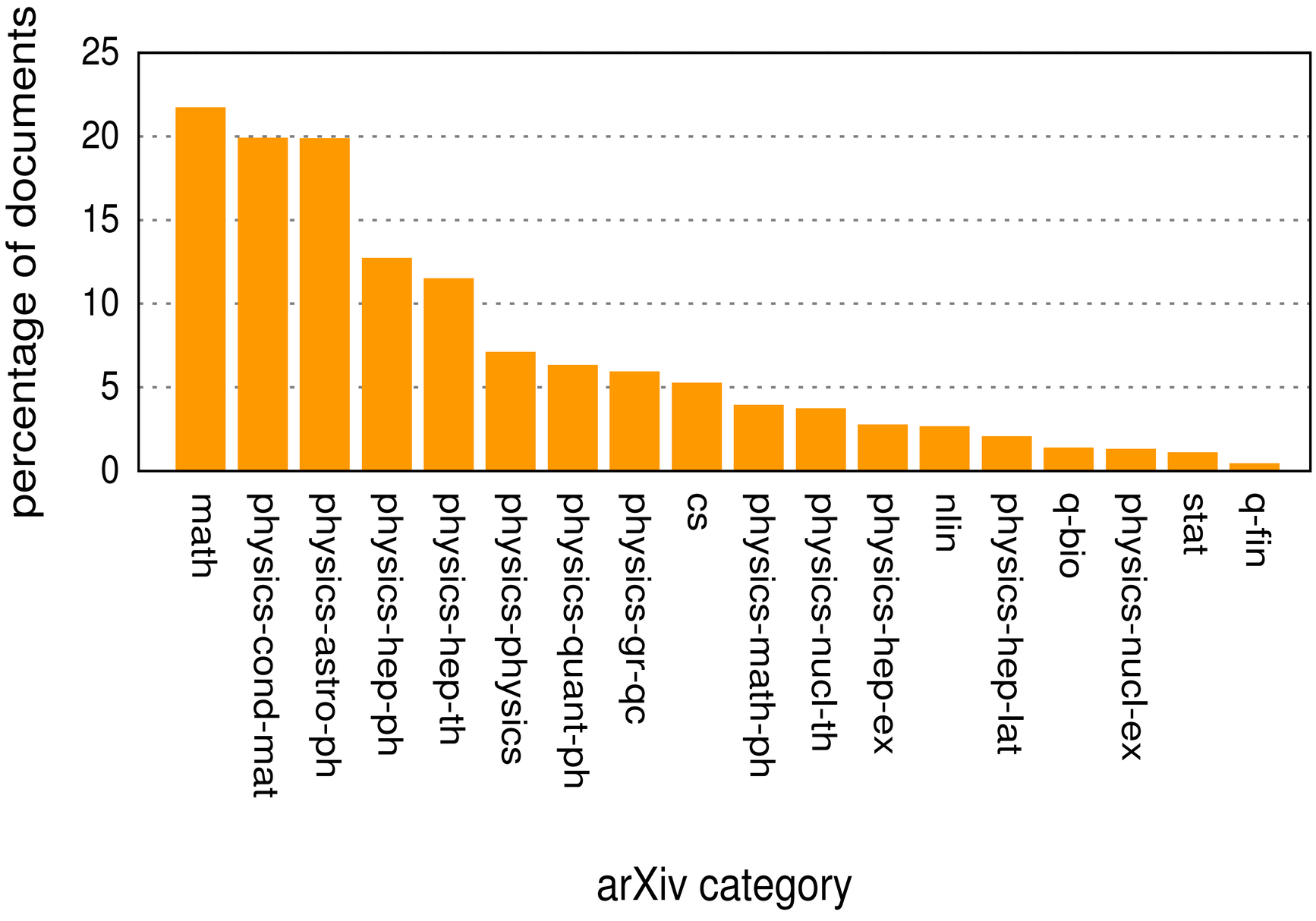
  Zipf's law for tags and distribution of distinct tags per document

- **Summary and outlook**

What? Why? How?

What data we use?

- **Abstracts and titles from arxiv.org (1991 - 03.2012)**
- **0.7 million documents from various fields of science**



What we do?

Example – arXiv id: 0704.2167, disciplines: math, stats

On the Computational Complexity of MCMC-based Estimators in Large Samples

In this paper we examine the implications of the statistical large sample theory for the computational complexity of Bayesian and quasi-Bayesian estimation carried out using Metropolis random walks. Our analysis is motivated by the Laplace-Bernstein-Von Mises central limit theorem, which states that in large samples the posterior or quasi-posterior approaches a normal density. Using the conditions required for the central limit theorem to hold, we establish polynomial bounds on the computational complexity of general Metropolis random walks methods in large samples. Our analysis covers cases where the underlying log-likelihood or extremum criterion function is possibly non-concave, discontinuous, and with increasing parameter dimension. However, the central limit theorem restricts the deviations from continuity and log-concavity of the log-likelihood or extremum criterion function in a very specific manner.

Under minimal assumptions required for the central limit theorem to hold under the increasing parameter dimension, we show that the Metropolis algorithm is theoretically efficient even for the canonical Gaussian walk which is studied in detail. Specifically, we show that the running time of the algorithm in large samples is bounded in probability by a polynomial in the parameter dimension d, and, in particular, is of stochastic order d^2 in the leading cases after the burn-in period. We then give applications to exponential families, curved exponential families, and Z-estimation of increasing dimension.

Tags from dictionary based on Wikipedia (WIKI)

approaching normal, bayesian estimate, central limit theorem, computational complexity, criterion function, exponential families, large sample, large sample theory, leading case, limit theorem, log concave, log likelihood, Metropolis algorithm, non concave, random walk, run time, sampling theory, stochastic order, von Mises

Tags from dictionary based on noun phrases found in the whole corpus (NP)

based estimates, bayesian estimates, central limit, central limit theorem, computation complexity, criterion function, exponential families, increasing dimension, large sample, large sample theory, limit theorem, log concave, log likelihood, metropolis algorithm, minimal assumption, normal densities, polynomial bounds, possible non, random walk, run time, sampling theory, specific manner, stochastic order, underlying log, von Mises

Be patient – the details of the method follow in two slides...

Why we do it?

- **To have better features (going beyond bag of words representation) for ML tasks such as document similarity, clustering, topic modelling, etc.**
- **To compare noun phrases based method (NP) and Wikipedia approach (WIKI)**
  - **Wikipedia is a general purpose lexicon, is it enough for scientific texts?**
  - **How the terms coverage depends on scientific discipline?**
  - **Tagging by team of experts infeasible (no "ground truth"), hence comparison of independent WIKI & NP methods yields valuable insight**
- **To examine statistical properties of dictionary tags**

How we do it?

- **Generate dictionary**
  - **WIKI – take all multiword entries in Wikipedia**
  - **NP – take all noun-phrases detected by OpenNLP, which occur more than 3 times**

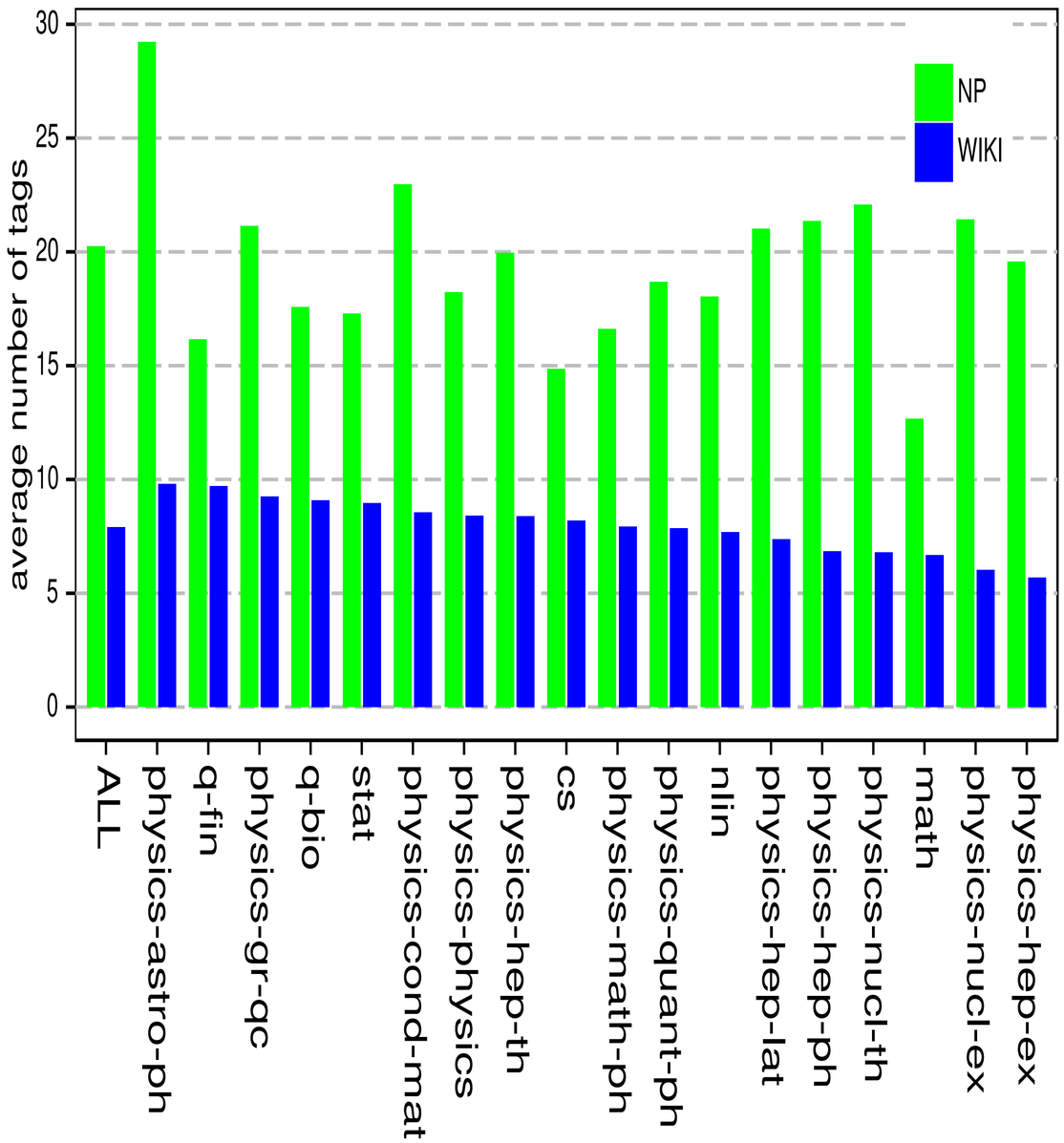
- **Clean dictionary using heuristics**
  - **Remove initial and final word, if they belong to stopwords**
  - **Remove all entries that contain one word**
  - **Remove all the entries that contain stopwords [Rose et al, 2010]**

- **Mark each paper using obtained dictionary**
  - **Use Porter stemming to capture different grammatical forms**

Comparison of the WIKI and NP Methods

Comparison – number of tags per document (1)

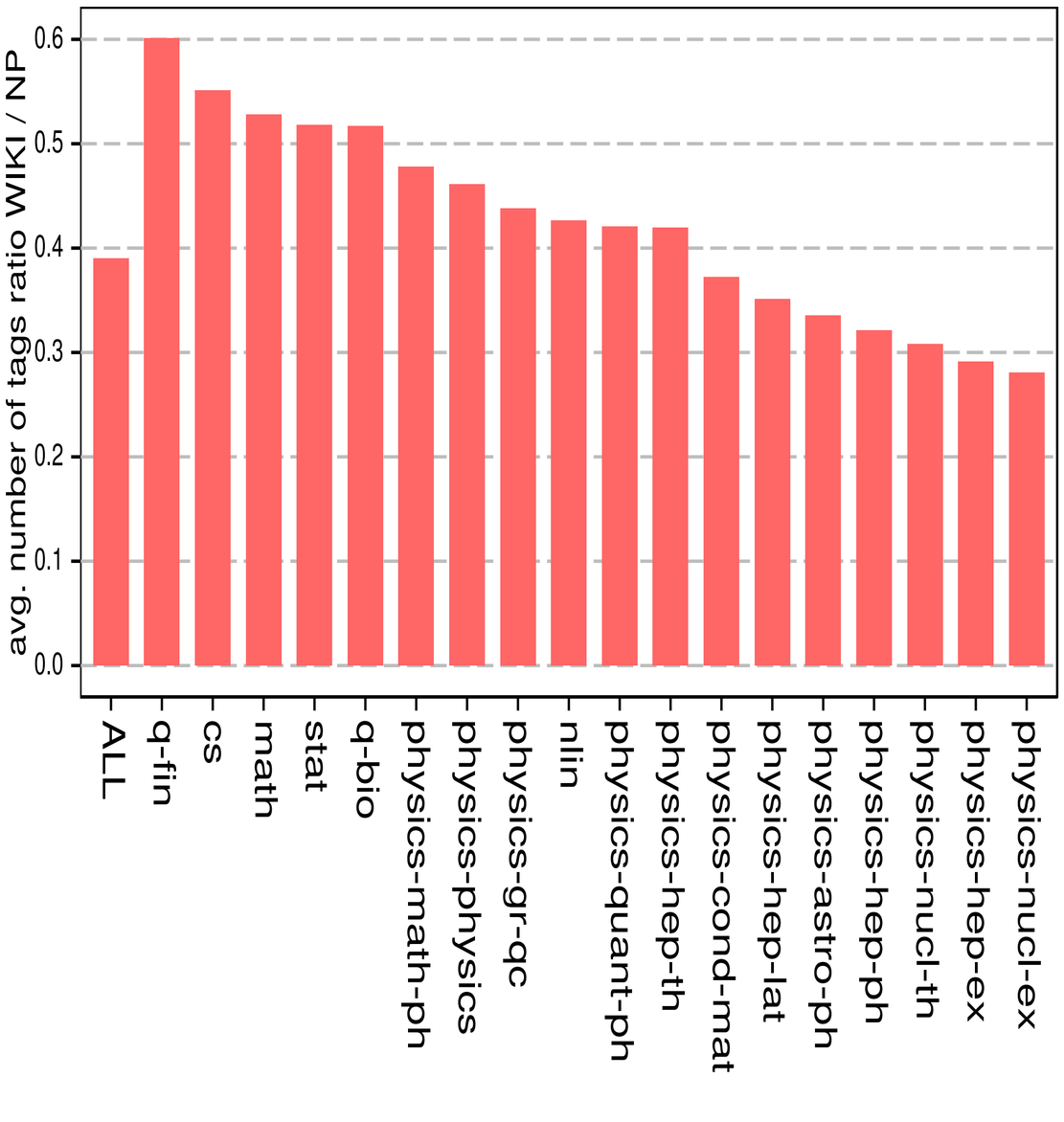
Average number of tags per doc. from NP & WIKI methods



- **Average number of tags per document strongly depends on discipline**
- **There is almost no correlation between WIKI and NP across disciplines (high avg. number of tags in WIKI does not imply high avg. number of tags in NP)**
- **Quantified by correlation coefficient ρ=0.13**

Comparison – number of tags per document (2)

Ratio of
average number of tags per doc.
from NP & WIKI methods



- **Average number of WIKI tags is within 30-60% of the NP result**
- **Higher ratios for most "everyday fields" (cs, q-fin)**
- **Lower ratios for exotic fields (nucl-ex, hep-ex)**

Comparison – category math

Detects additional tags related to the NP. Combining NP + NER could improve the situation.

A few uninformative tags are present (imperfect filtering)

Top WIKI	Top NP	Top WIKI-only	Top NP-only
Lie algebra	Lie algebra	Calabi Yau	give rise
differential equation	differential equation	Navier Stokes	higher order
moduli space	moduli space	point of view	initial data
lower bound	lower bound	non negative	infinitely many
field theory	field theory	Cohen Macaulay	new proof
finite dimensional	finite dimensional	algebraically closed	over field
sufficient condition	sufficient condition	degrees of freedom	value problem
upper bound	upper bound	self dual	large class
Lie group	Lie group	Gromov Witten	time dependence
two dimensional	two dimensional	answered question	mapping class

Top tags are identical for the WIKI and NP case

A few incomplete tags are detected by the NP (imperfect POS tagger)

Comparison – category physics-nucl-ex

Accident – Au Au links to auction portal description in Wikipedia

NP detects many high rank tags not present in WIKI, to specific to be described in Wikipedia

Top WIKI	Top NP	Top WIKI-only	Top NP-only
cross section	cross section	equation of state	heavy ion
Au Au	heavy ion	center of mass	Au collisions
heavy ion collision	Au Au	order of magnitude	ion collision
form factor	Au collisions	degrees of freedom	Au Au collision
beta decay	ion collision	ultra relativistic	transversal momentum
elliptic flow	Au Au collision	Drell Yan	200 GeV
high energies	heavy ion collision	time of flight	relativistic heavy
experimental data	transversal momentum	presented first	relativistic heavy ions
charged particle	200 GeV	long lived	low energies
nuclear matter	form factor	national laboratory	Pb Pb

Top tags are different for NP and WIKI

Comparison – $C_{WIKI}(r)$ and $C_{NP}(r)$

- **The previous slides suggest that first *r* tags can be either identical or different for a particular discipline**

- **Let's quantify it by counting the percentage of unique tags up to rank *r* for each discipline in WIKI/NP methods**

$T_{\mathrm{WIKI}}(r)$ **– set of WIKI tags up to rank *r***

$T_{\mathrm{NP}}(\infty)$ **– set of all NP tags**

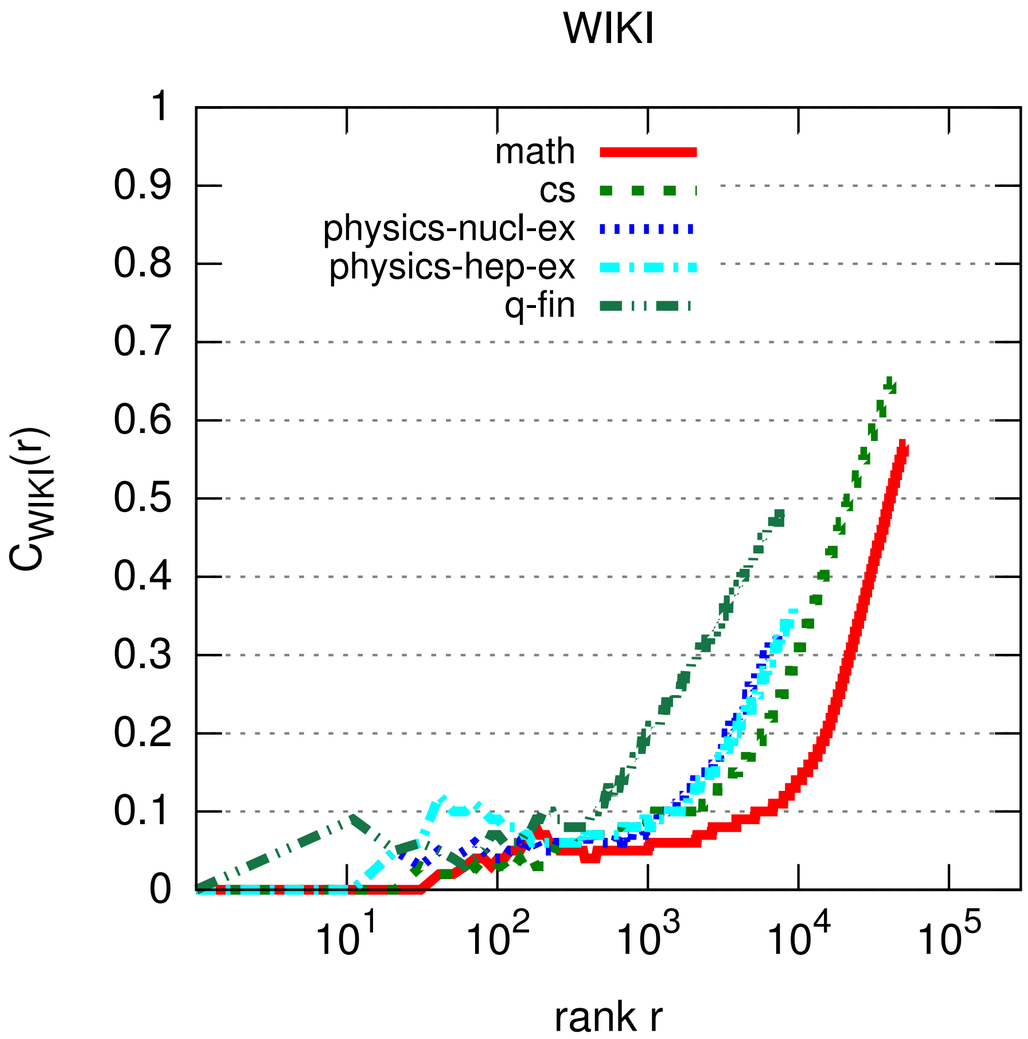
$$C_{\mathrm{WIKI}}(r) = \frac{\#(T_{\mathrm{WIKI}}(r) \setminus T_{\mathrm{NP}}(\infty))}{r}$$

Number of WIKI tags up to rank **r** NOT included in all NP tags

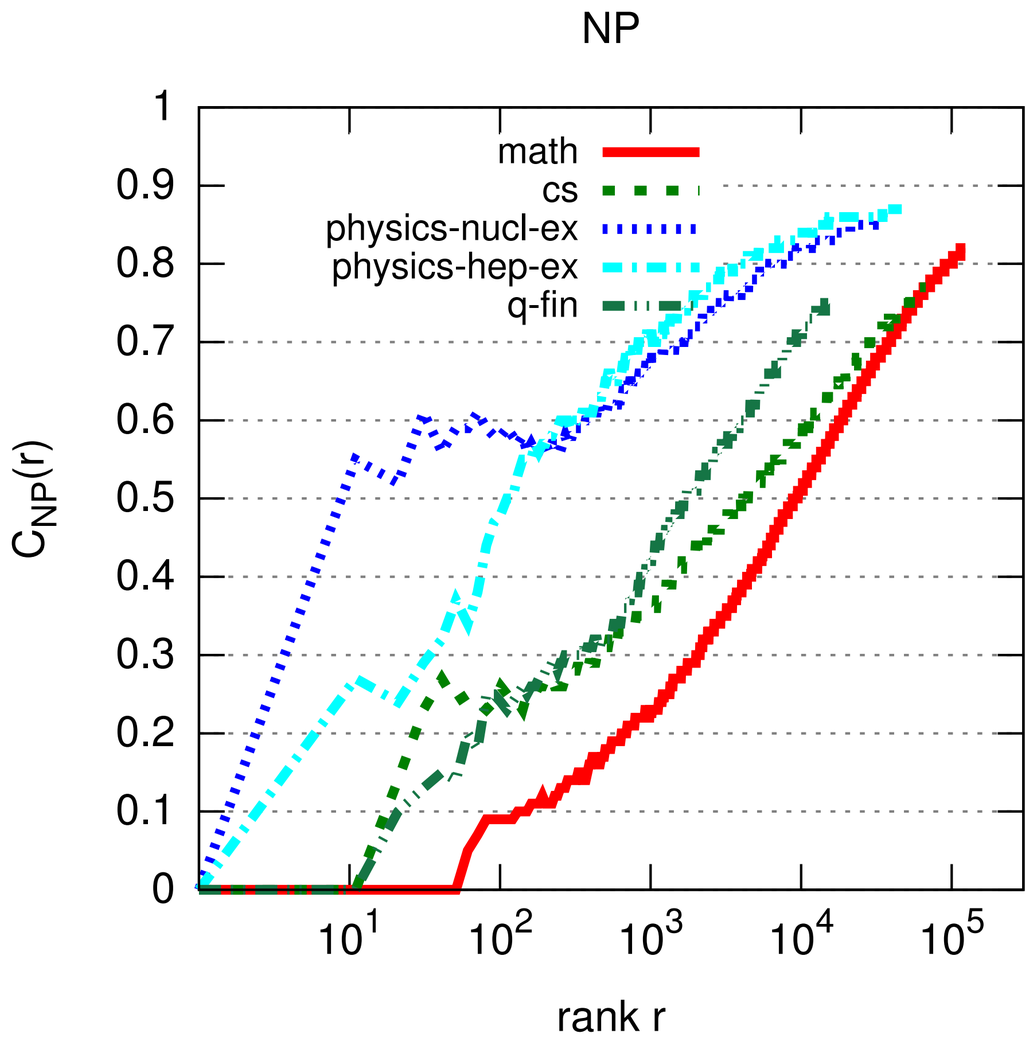
Divide by rank **r** to normalize

$C_{NP}(r)$ – defined in the analogous way

Comparison – $C_{WIKI}(r)$ and $C_{NP}(r)$



- **Only 10% of the WIKI tags not detected by the NP up to high ranks ~ 1000**



- **The percentage of unique NP tags strongly depends on discipline**
- **The more exotic the discipline the faster is the increase of $c_{NP}(r)$**

Statistical Properties of Tags

Statistics – Zipf's law

- **Zipf's law for words**

  Word frequency *f* as a function of its rank *r* exhibits power-law behaviour

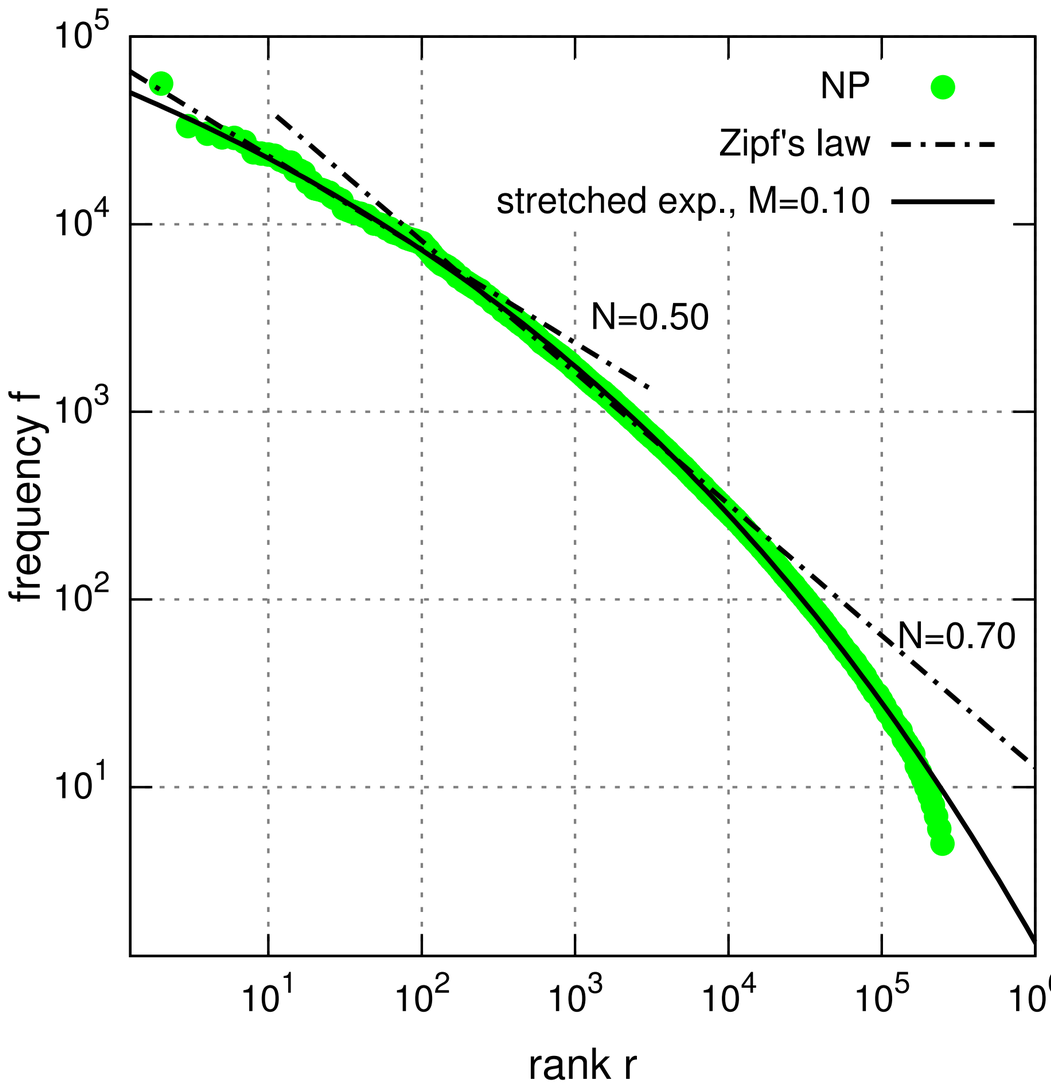
$$f(r; A, N) = A\, r^{-N}$$

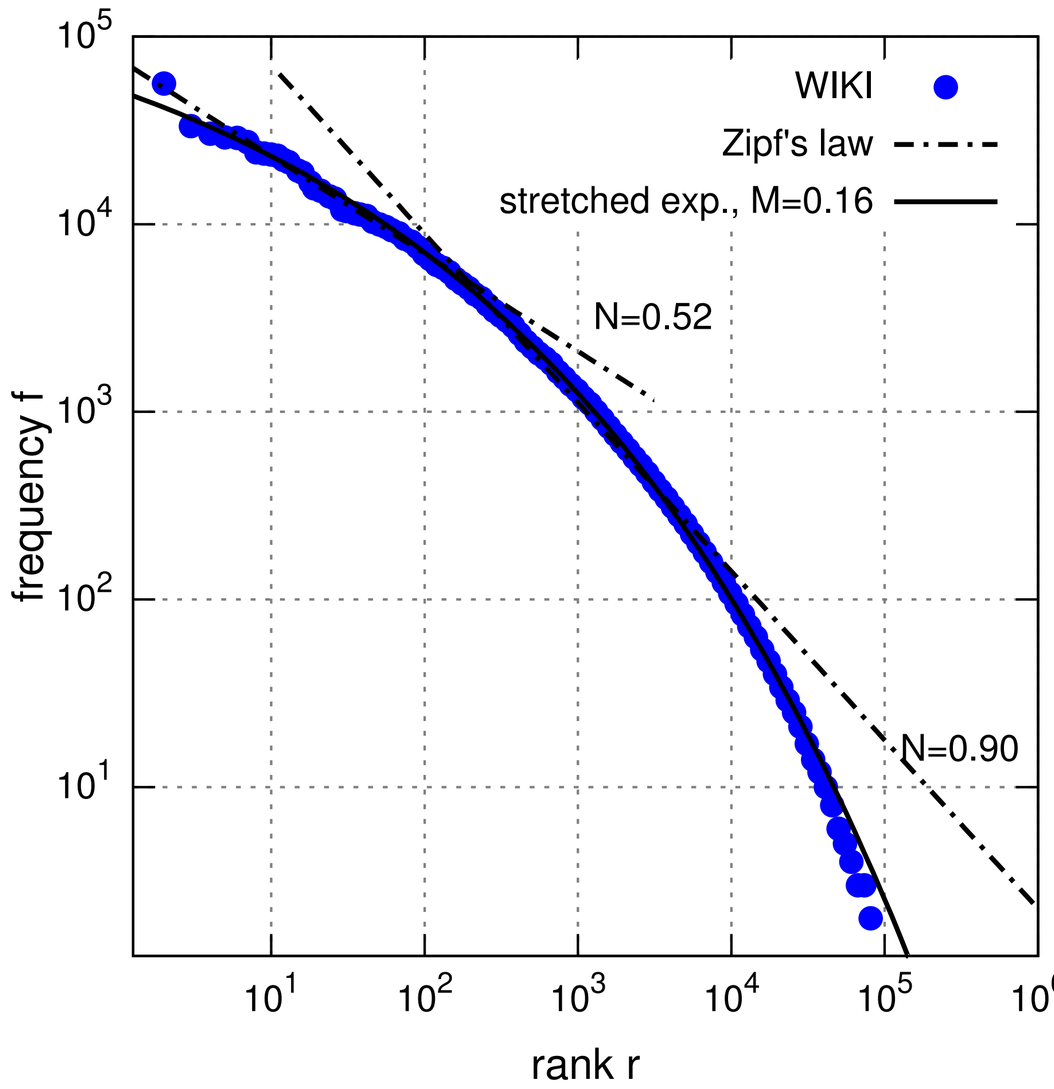
- **Is Zipf's law valid for discussed dictionary tags?**
- **Are there qualitative differences between WIKI & NP?**

Statistics – rank-frequency curves for tags

- **Only approximately follow Zipf's Law**
- **Better described by the stretched exp. [Laherrère, 1998]**

$$f(r; C, D, M) = C \exp\left(-D\, r^{M}\right)$$

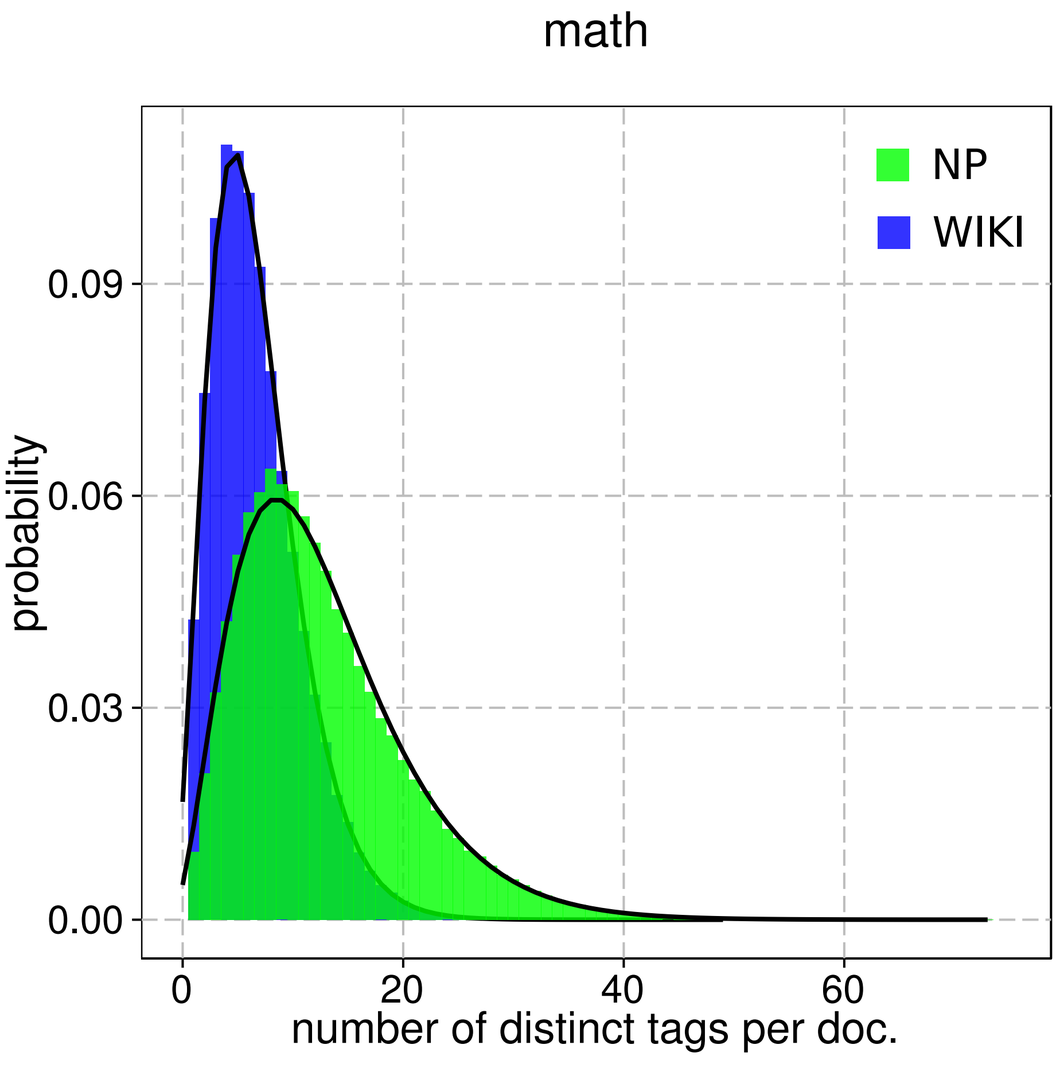


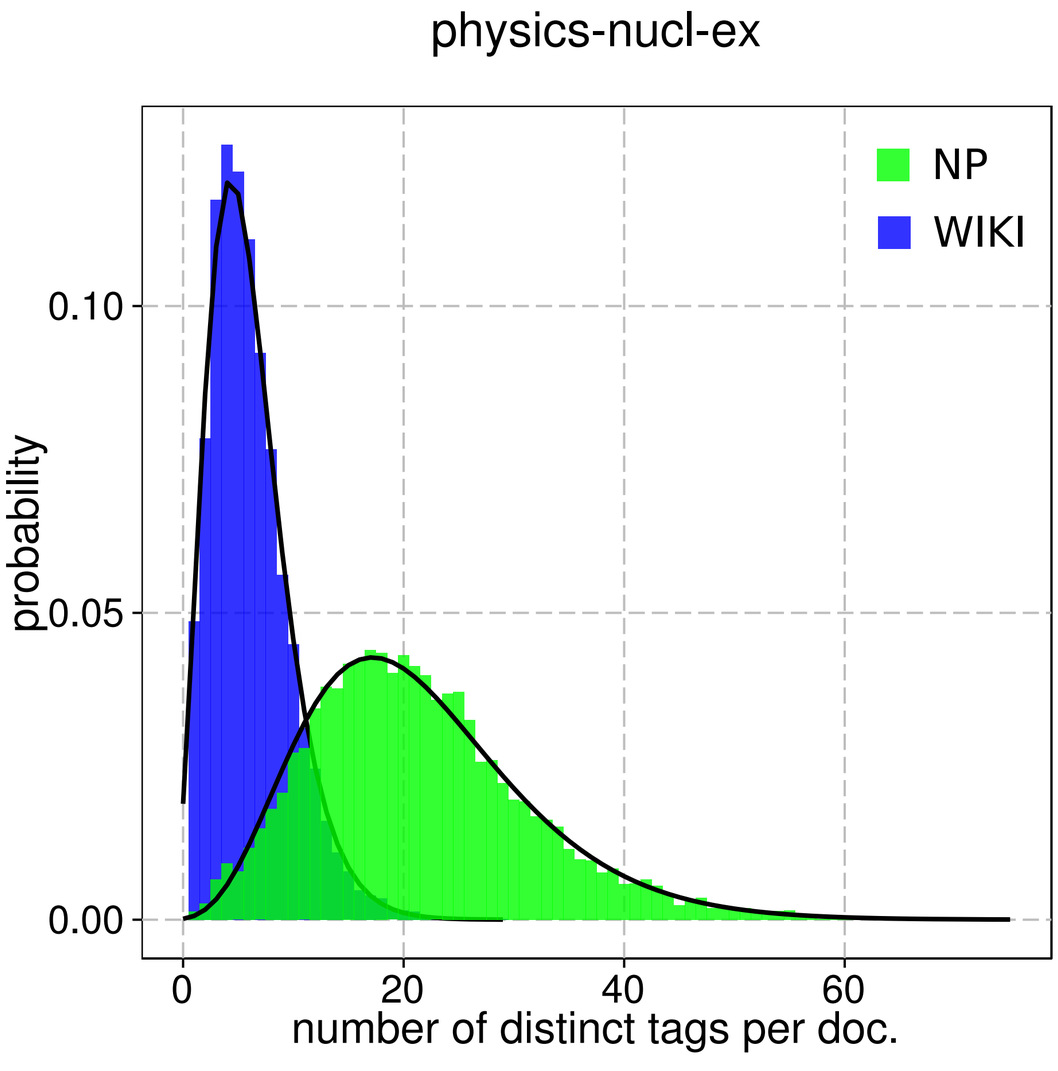


Statistics – distribution of #tags per document

- **Distribution of number of distinct tags per document can be well described with negative binomial model**

$$\mathrm{Prob}(k; P, R) = \binom{k+R-1}{k} P^R (1-P)^k,$$





Summary and Outlook

Summary and outlook

- **Comparison of tagging by the WIKI & NP methods**
  - **NP yields 2-3 times more tags than WIKI**
  - **WIKI coverage is better for more "everyday" fields such as cs or finance, worse for exotic ones, e.g., nuclear or HEP physics**
  - **NP sometimes yields "broken phrases" due to NLP tools imperfections**
  - **WIKI is much better at detecting tags related to surnames**
  - **Both WIKI & NP generated certain fraction of uninformative tags. This could be improved by tweaking filtering phase**
- **Statistical properties of generated tags**
  - **WIKI & NP tags have qualitatively identical statistical properties**
  - **Rank-frequency curve can be approximated by stretched exponential**
  - **Number of tags per doc. follows negative binomial model**
- **Outlook**
  - **Tweak the approach (e.g., filtering) & assess it on ML tasks**

Acknowledgements

This research was carried out with the support of the ”HPC Infrastructure for Grand Challenges of Science and Engineering (POWIEW)” Project, co-financed by the European Regional Development Fund under the Innovative Economy Operational Programme

POWIEW/ HPC Infrastructure for Grand Challenges of Science and Engineering

INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND

Thank you!

Questions?