

Turn Email into Data with Deep Learning

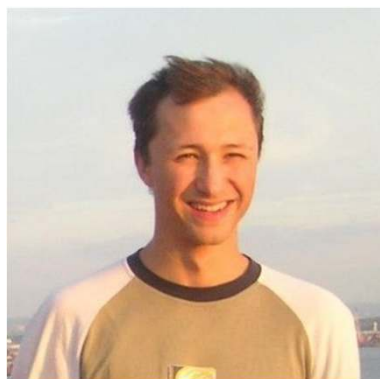
Plus Other Industry Tasks with Gensim Topic Modeling



<http://rare-technologies.com/>

Lev Konstantinovskiy

About



Lev Konstantinovskiy

@teagermylk

lev@rare-technologies.com

NLP consultant at RaRe Technologies

Community manager of Gensim Open Source Project

Background in Financial Trading and Mathematics

We are a ML consulting organisation



DATASIFT

amazon



H E A R S T



elevate

DynAdmic

>Eva
Expert Virtual Agent

harvest.ai

**SPORTS
AUTHORITY.**

r/ally

People▲Ticker.



Topic Modelling

Using Gensim

Client: publicly traded mass media company

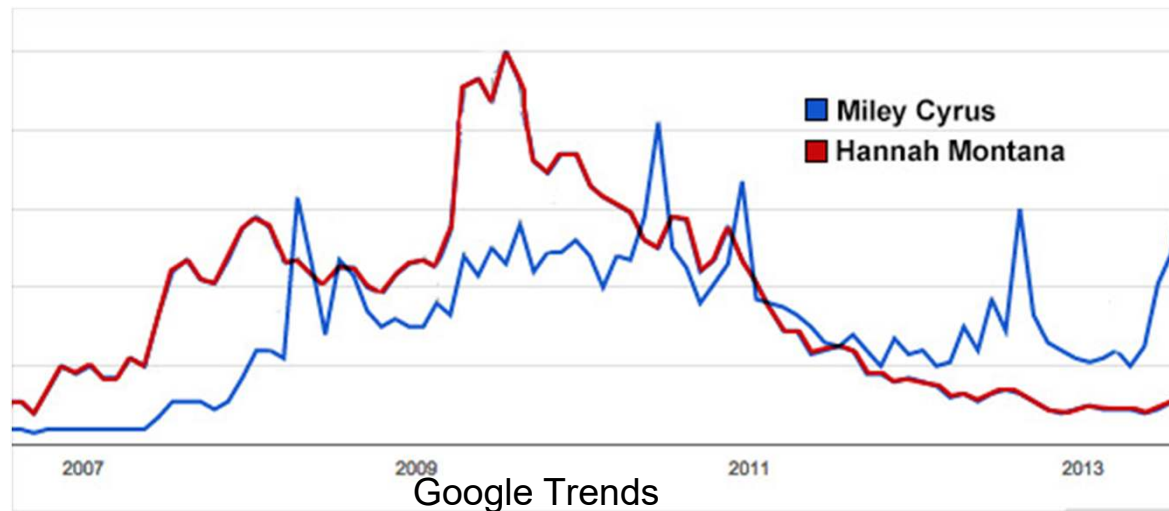
Business problem: How is the CELEBRITY content driving revenue this month?

Technical problem: search.

Find all CELEBRITY articles

Which keywords to search for?

Technical problem: find all CELEBRITY articles
Which keyword to search for?



Remove “Hannah Montana” keyword in 2011.

Add “Miley

Cyrus” back in 2012
Maintaining keywords is expensive

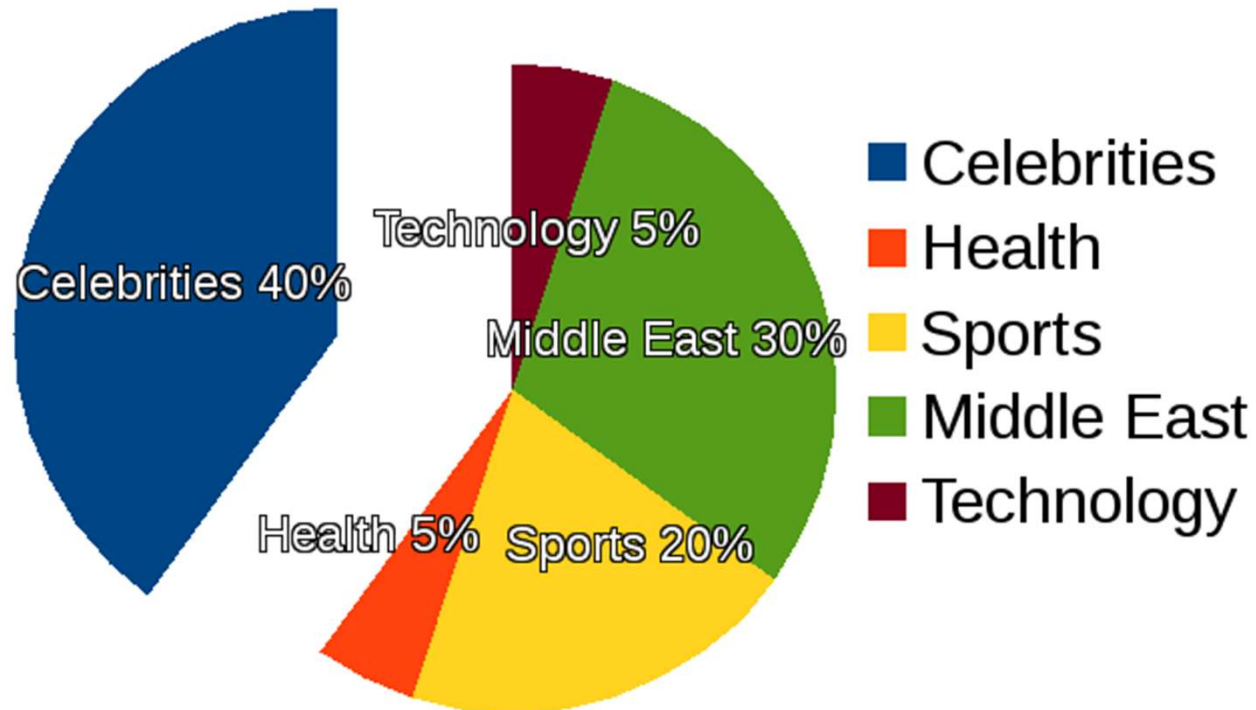
Better than keywords

“You shall know a word by the company it keeps”

An algorithm can group together words that appear together.

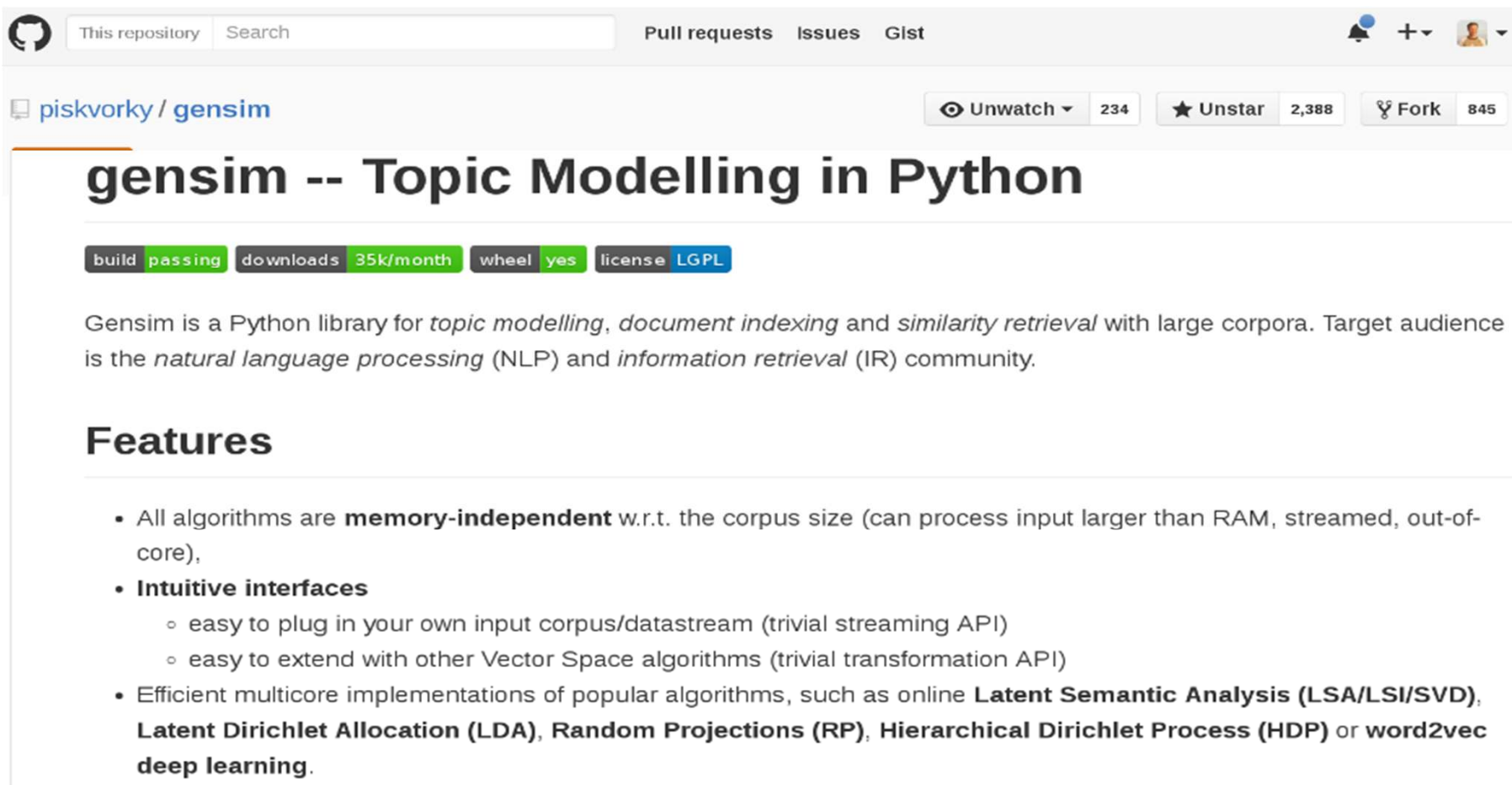
We call these groups of words **Topics**.

Solution: Search by Topic



**Topic Model needs no manual labor
compared to keywords, taxonomy or**

Gensim open-source package



The screenshot shows the GitHub repository for Gensim, a Python library for topic modelling. The repository is owned by piskvorky. The page displays the repository name, a search bar, and navigation links for Pull requests, Issues, and Gist. The repository has 234 watchers, 2,388 stars, and 845 forks. The main heading is "gensim -- Topic Modelling in Python". Below the heading, there are badges for build status (passing), downloads (35k/month), wheel support (yes), and license (LGPL). The description states that Gensim is a Python library for topic modelling, document indexing, and similarity retrieval with large corpora, targeting the natural language processing (NLP) and information retrieval (IR) community. The Features section lists several key capabilities: memory-independent algorithms, intuitive interfaces for plugging in corpora and extending with other algorithms, and efficient multicore implementations of popular algorithms like LSA/LSI/SVD, LDA, RP, HDP, and word2vec deep learning.

This repository Search Pull requests Issues Gist

piskvorky / gensim Unwatch 234 Unstar 2,388 Fork 845

gensim -- Topic Modelling in Python

build passing downloads 35k/month wheel yes license LGPL

Gensim is a Python library for *topic modelling*, *document indexing* and *similarity retrieval* with large corpora. Target audience is the *natural language processing* (NLP) and *information retrieval* (IR) community.

Features

- All algorithms are **memory-independent** w.r.t. the corpus size (can process input larger than RAM, streamed, out-of-core),
- **Intuitive interfaces**
 - easy to plug in your own input corpus/datastream (trivial streaming API)
 - easy to extend with other Vector Space algorithms (trivial transformation API)
- Efficient multicore implementations of popular algorithms, such as online **Latent Semantic Analysis (LSA/LSI/SVD)**, **Latent Dirichlet Allocation (LDA)**, **Random Projections (RP)**, **Hierarchical Dirichlet Process (HDP)** or **word2vec deep learning**.

Gensim Open Source Package



4 Corporate



- Numerous Industry Adopters
- 140 Code contributors, 3000 Github stars
- 200 Messages per month on the mailing list
- 100 People chatting on Gitter
- 380 Academic citations

The Gensim algorithm block is nice, but...

increasing resource efficiency is nicer.

How to apply it to my domain? (media, HR, legal etc)

How to have a view of my business?

How to integrate with my analytics suite?

How to make it robust?

The business value is in the application.

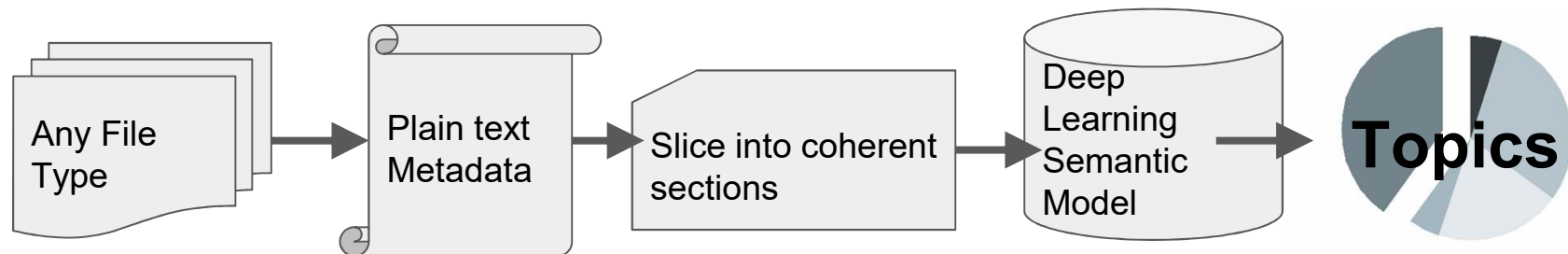
ScaleText

User-friendly Topic Modelling Solution

The business value is in the application

ScaleText

User-friendly topic modelling solution



Specific modules for media, HR, legal

Another way to drive business value

Not just Topic Modelling...

Information Extraction

*Turn unstructured text into structured tables with
deep learning*

Industry setting: wood trucks moving across Canada



Business problem: extract data from truck reports

Content: *A truck of type “Englewood” owned by ForestCo left Cold Stream forest on 26 August for the mill in Enderby carrying 140 logs of wood at the rate of \$10k.*

In an email it looks like this:

ENGLEWOOD	140	26/08	Cold Stream/Enderby	10k	ForestCo
-----------	-----	-------	---------------------	-----	----------

Problem: Constantly changing *100* formats

In an email it looks like this:

ENGLEWOOD 140 26/08 Cold Stream/Enderby 10k ForestCo

Sometimes like this:

26/08 ENGLEWOOD ForestCo 140 Cold Stream to Enderby at 10k

Or even like this:

ForestCo Cold Stream==Enderby 26/08 ENGLEWOOD 140 - 10k

Would you like to maintain 100 changing regexes?

Model: Deep bi-directional LSTM network

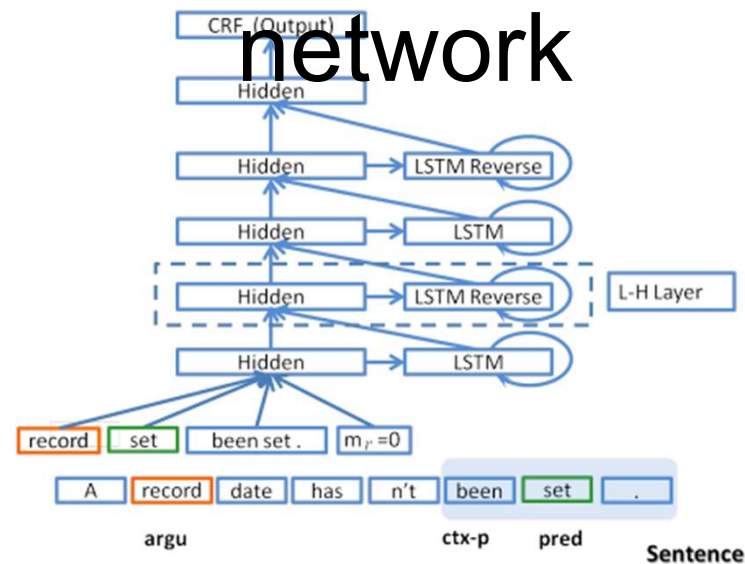


Figure 2: DB-LSTM network. Shadow part denote the predicate context within length 1.

[End-to-end learning of semantic role labeling using recurrent neural networks](#) Zhou & Xu
International joint conference on Natural Language Processing, 2015

Task: Character-level annotation

L244:ENGLEWOOD 140 26/08 Cold Stream/Enderby 10k ForestCo
Pred:vvvvvvvvv-----qqq---tt-tt--11111111111-uuuuuuu-----rrr-----cccccccc

Labels: [u]nloading, [l]oading, [c]ompany,
[t]ime, [r]ate, [v]ehicle, [-]junk_field, [q]uantity

Deep Learning Tricks

Trick: generate canned data to supplement manual annotations

Result: increase accuracy by 30%

Model Performance

Business value: no manual labor to maintain 100 regexes anymore.

Performance metric: only exact match in all characters is valuable to the client.

When confidence is low - ask a human.

Human in the loop alerting on: 5% lines

Accuracy achieved: 96% of lines match exactly on every character.

Business metrics more important than algos and code

- Algorithms don't know how to drive value
- Open source software is only a part of the solution
- Achieving business goals requires an entire production class ML application

We do theoretical papers, practical software...

but most of all we believe in executing on

Business metrics.

Open source Python NLP eco-system



RARE Training

- customized, interactive [corporate training](#) hosted on-site for technical teams of 5-15 developers, engineers, analysts and data scientists
- 2-day intensives include [Tensorflow Training](#), [Python Best Practices](#) and [Practical Machine Learning](#), and 1-day intensive [Topic Modelling](#)

industry-leading instructors



**RNDr. Radim
Řehůřek, Ph.D.**



**Gordon Mohr,
BA in CS & Econ**

for more information email
training@rare-technologies.com

Q&A

Lev Konstantinovskiy

If you need help with solving your business problems or Training

lev@rare-technologies.com

Twitter @teagermylk

