Text Mining For Definitions in the arXiv

Luis Berlioz
lab232@pitt.edu

Formal Methods in Mathematics / Lean Together 2020

January 10, 2020

Objectives and Outline

Objective

Create a machine learning system that can find the definitions and the terms being defined in large collections of mathematical texts.

The problem is broken down into two parts:

The Classifier: Tells if a given paragraph is a definition or not

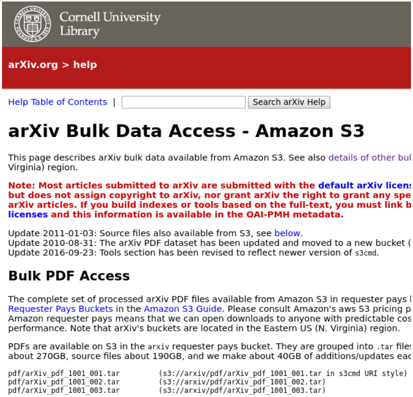
A Named Entity Recognition system: given a definition, returns the term that is being defined (definiendum).

For each part I will describe how to:

- Get and process the relevant data.
- Train the machine learning system.
- Take a look at the results.

arXiv Website Bulk Download

All the LaTeX source files can be downloaded from an Amazon S3 bucket



arXiv Bulk Data Access - Amazon S3

This page describes arXiv bulk data available from Amazon S3. See also details of other bul
Virginia) region.

**Note: Most articles submitted to arXiv are submitted with the default arXiv licens
but does not assign copyright to arXiv, nor grant arXiv the right to grant any spe
arXiv articles. If you build indexes or tools based on the full-text, you must link b
licenses and this information is available in the OAI-PMH metadata.**

Update 2011-01-03: Source files also available from S3, see below.
Update 2010-08-31: The arXiv PDF dataset has been updated and moved to a new bucket (
Update 2016-09-23: Tools section has been revised to reflect newer version of `s3cmd`.

Bulk PDF Access

The complete set of processed arXiv PDF files available from Amazon S3 in requester pays l
Requester Pays Buckets in the Amazon S3 Guide. Please consult Amazon's aws S3 pricing p
Amazon requester pays means that we can open downloads to anyone with predictable cos
performance. Note that arXiv's buckets are located in the Eastern US (N. Virginia) region.

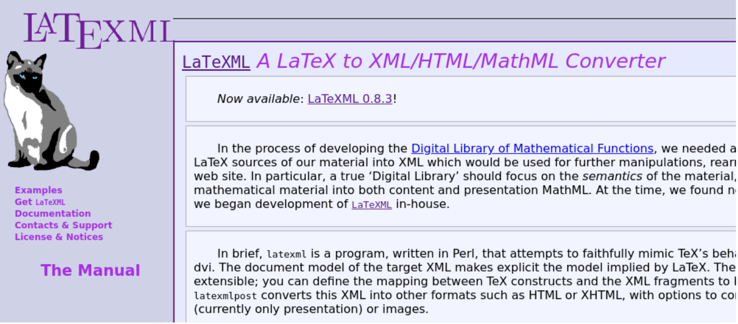
PDFs are available on S3 in the `arxiv` requester pays bucket. They are grouped into `.tar` files
about 270GB, source files about 190GB, and we make about 40GB of additions/updates eac

```
pdf/arXiv_pdf_1001_001.tar        (s3://arxiv/pdf/arXiv_pdf_1001_001.tar in s3cmd URI style)
pdf/arXiv_pdf_1001_002.tar        (s3://arxiv/pdf/arXiv_pdf_1001_002.tar)
pdf/arXiv_pdf_1001_003.tar        (s3://arxiv/pdf/arXiv_pdf_1001_003.tar)
```

- About 1 Terabyte of .tar files.
- Each .tar file is about 500 Megabytes.
- Download without affecting the website's traffic.
- LaTeX source is converted to a more structured format.

LaTeXML

Process each article to get a more structured format



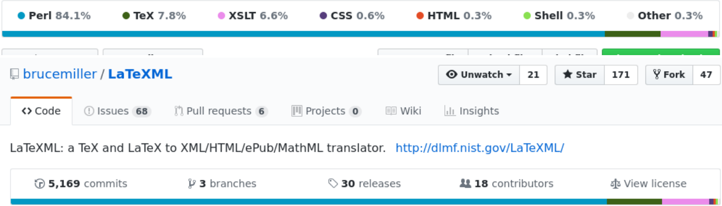
LaTeXML

LaTeXML *A LaTeX to XML/HTML/MathML Converter*

Now available: LaTeXML 0.8.3!

In the process of developing the Digital Library of Mathematical Functions, we needed a LaTeX sources of our material into XML which would be used for further manipulations, rearr web site. In particular, a true 'Digital Library' should focus on the *semantics* of the material, mathematical material into both content and presentation MathML. At the time, we found n we began development of LaTeXML in-house.

In brief, `latexml` is a program, written in Perl, that attempts to faithfully mimic TeX's beha dvi. The document model of the target XML makes explicit the model implied by LaTeX. The extensible; you can define the mapping between TeX constructs and the XML fragments to l `latexmlpost` converts this XML into other formats such as HTML or XHTML, with options to co (currently only presentation) or images.

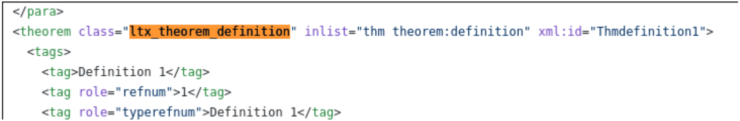


LaTeXML: a TeX and LaTeX to XML/HTML/ePub/MathML translator. http://dlmf.nist.gov/LaTeXML/

credit: Bruce Miller, `www.nist.gov/people/bruce-r-miller`

Obtaining and Classifying Definitions

- Sometimes the author of an article uses a LaTeX macro to label a definition. These are our positive labels:

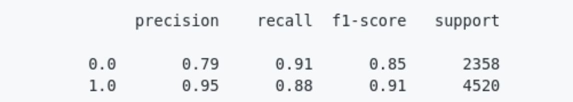


```
</para>
<theorem class="ltx_theorem_definition" inlist="thm theorem:definition" xml:id="Thmdefinition1">
  <tags>
    <tag>Definition 1</tag>
    <tag role="refnum">1</tag>
    <tag role="typerefnum">Definition 1</tag>
```

- To get examples of non-definitions, we pick paragraphs at random and assume they are not definitions.
- This has the drawback that some of the non-definitions are wrong.
- There are 1,707 articles in 2015 math.AG, we go from 5,229 labeled definitions to 71,067 “probable” definitions.

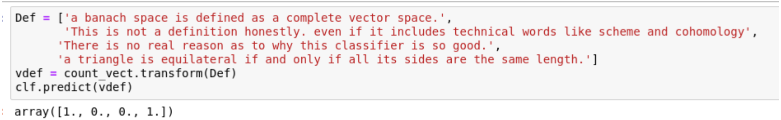
Some Classification Results

- Results using SVC (Support Vector Classifier) in **scikit-learn**



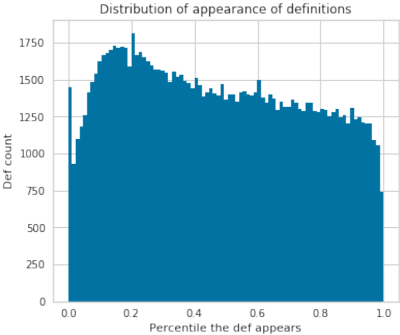
```
             precision    recall  f1-score   support

        0.0       0.79      0.91      0.85      2358
        1.0       0.95      0.88      0.91      4520
```

- Sanity check:



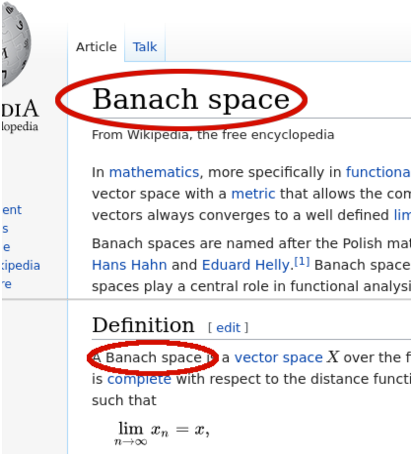
```
Def = ['a banach space is defined as a complete vector space.',
        'This is not a definition honestly. even if it includes technical words like scheme and cohomology',
       'There is no real reason as to why this classifier is so good.',
       'a triangle is equilateral if and only if all its sides are the same length.']
vdef = count_vect.transform(Def)
clf.predict(vdef)
```

```
array([1., 0., 0., 1.])
```



Extracting the Definienda

Obtaining the data for Named Entity Recognition system



Banach space

From Wikipedia, the free encyclopedia

In mathematics, more specifically in functiona
vector space with a metric that allows the con
vectors always converges to a well defined lin

Banach spaces are named after the Polish mat
Hans Hahn and Eduard Helly.[1] Banach space
spaces play a central role in functional analysi

Definition [edit]

A Banach space is a vector space X over the f
is complete with respect to the distance functi
such that

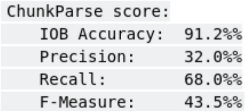
$$\lim_{n\to\infty} x_n = x,$$

- Go through every of wikipedia article looking for a Definition section that contains the title.
- We obtain a pair: (**Definienda**, Definition).
- Just 5,321 matches out of almost 6 million articles.
- Other websites/datasets:
  - All types of wikis, e.g. ProofWiki, GroupProps (500)
  - The Stacks project (3,000)
  - PlanetMath (1,500)

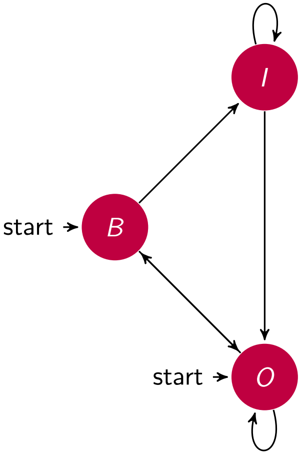
Training and Evaluating the NER System

Results of the IOB parser using the ChunkParserI method in the **nltk** library

Input		Output
Token	POS	NER
We	PRP	O
define	VBP	O
a	DT	O
Banach	NNP	B–DFNDUM
space	NN	I–DFNDUM
as	IN	O
a	DT	O
complete	JJ	O
vector	NN	O
space	NN	O



```
ChunkParse score:
    IOB Accuracy:  91.2%%
    Precision:     32.0%%
    Recall:        68.0%%
    F-Measure:     43.5%%
```



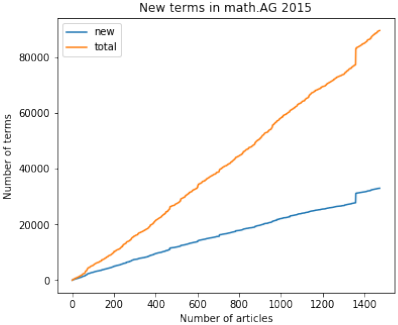
Some definitions found in the 2015 math.DG articles:

Ex. Things we like

An induced generalized Kähler structure on _inline_math_ is a Lie algebraic generalized Kähler structure with _inline_math_. It is a canonical generalized Kähler structure if _inline_math_.

Ex. Things we don't like

Suppose _inline_math_ is a vector space. The only connection on the graded manifold _inline_math_ is the trivial connection.



Conclusions and Future Work

- We think that we have collected enough evidence to believe that a robust collector of definitions is possible.
- A lot of interesting work ahead:
  - Organize the definitions in a *dependency tree structure.*
  - Produce word embedding with math tokens
    (e.g. where *Banach space* is just one token).
  - Apply disambiguation and polysemy detection techniques.

Text Mining For Definitions in the arXiv

Luis Berlioz
lab232@pitt.edu

Formal Methods in Mathematics / Lean Together 2020

January 10, 2020