Agenda

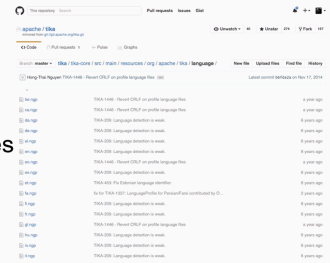
- Why Language Identification / Detection
- Representing and Standardized Language Codes (ISO-619)
- Language Identification Approaches
  - By Tagging / Humans
  - Automatic
    - Word Class
    - N-grams
- Libraries for Language Detection

Why Language Detection?

- Content is increasingly created around the world
- Since 2012 most of the people who search for content are
  - Using their cell phone
  - Coming from countries outside of the USA and whose primarily local language, dialect, etc., is not necessarily English
  - Over 9000 languages remain in use today
    - CSCI 572 Search Engines & Information Retrieval Course http://sunset.usc.edu/classes/cs572_2015b/

Apache Tika leverages ISO-639

- Language Profiles stored as *.ngp files within the tika-core module
- 28 language profiles using the 2 char codes from ISO-639-1



Areas for Language Identification

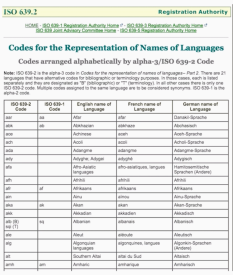
- Document Authoring – when creating the content, most often the operating system and content creator know the intended source language.
- Web Search Engines – need to understand how to tokenize and parse queries, and to index text and metadata in different languages.
- Content Detection and Analysis – for text, metadata extraction across languages and data.
- Web Browsing – Language matters for preferences, for websites, for domain names, URLs, etc.
- Sharing Research and Results – Increasingly research, business, and collaboration requires people around the world to work together. Being able to interpret and analyze that data requires understanding the language of the data and information.

Language Identification - defined

- "In natural language processing, language identification or language guessing is the problem of determining which natural language given content is in. Computational approaches to this problem view it as a special case of text categorization, solved with various statistical methods." – *Wikipedia* *https://en.wikipedia.org/wiki/Language_identification*

Codes for representation of languages

- ISO 639 – Language Codes
  - 6 different parts – 2 letter codes (*-1), 3 letter code (*-2), 3-letter codes for living and ancient languages (*-3), general usage of codes (*-4), families and groups (living and extinct) (*-5)
  - http://www.iso.org/iso/home/standards/language_codes.htm
    - Codes for the representation of names of languages-- Part 2: alpha-3 code
    - http://www.loc.gov/standards/iso639-2/
- 508 languages*
  - curl "http://www.loc.gov/standards/iso639-2/php/code_list.php" | grep \<tr | wc –l
    - 509 (subtract one for extra tr footer)
- 2 and 3 letter codes for languages, e.g., en, fr, etc.
- Last updated: 2014-03-18



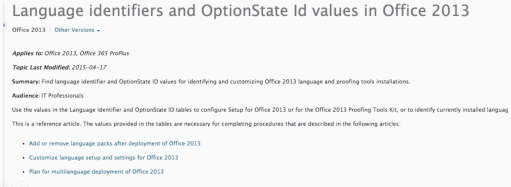
Tika ISO 639-1 support

- At time of writing Tika in Action, Tika supported 18 languages – has since grown to 28

- da—Danish
- de—German
- et—Estonian
- el—Greek
- en—English
- es—Spanish
- fi—Finnish
- fr—French
- hu—Hungarian
- is—Icelandic
- it—Italian
- nl—Dutch
- no—Norwegian
- pl—Polish
- pt—Portuguese
- ru—Russian
- sv—Swedish
- th—Thai

Microsoft Word has a Language Identifier

- Proprietary algorithm
- Present in Microsoft Office
  - Language Packs (defined custom codes)



Language Identification

- Basic Approaches
  - Human tagging of languages
  - Can be codified in e.g., HTML and other documents
- Function Words (R. Lins & P. Gonclaves 2004)
  - Indicative of Word Class
- N-gram based identification
  - Cavnar and Trenkle (1994), Dunning (1994)
    - Dunning, T. (1994) "Statistical Identification of Language". Technical Report MCCS 94-273, New Mexico State University, 1994.
  - Funny: Ted is now heavily involved in Big Data VP, Apache Incubator

Sometimes content authors tag HTML

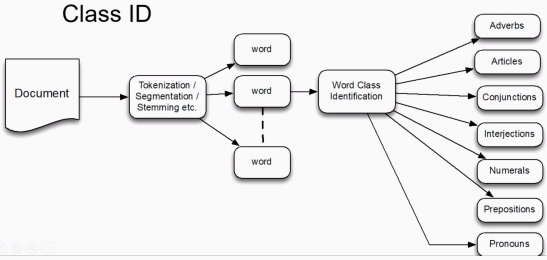
- XML:lang attribute
- HTML lang attribute

```
>Here is a list of part numbers: <span lang="zxx">9RUI34 8XO5I2 3TYY85</span>.</p>
```

- https://www.w3.org/International/questions/qa-no-language
- Of course this is best, but doesn't scale
  - Beyond that, sometimes people forget to tag the language, etc.

Algorithm - Preprocessing

- Preprocessing
  - Tokenization, Stemming, Segmentation, Class ID

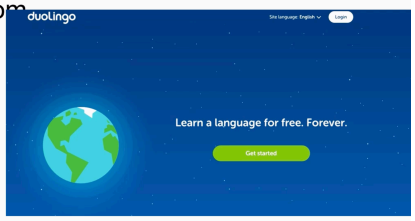


Algorithm – 1 Word Class

- 40% - Ratio of of percentage of words from text found in English class dictionary compared to the other languages
  - Threshhold computed based on looking at 4000 web pages and heuristically evaluating them
- 5% - Ratio of percentage of words from found in English dictionary for word class compared to total number of words in input text
- In English: If at least 5% of the input text was found in the English dictionary, and at least 40% of words found in all dictionaries were in the English dictionary, it's English

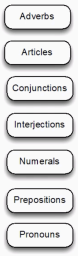
1. Human Tagging of Languages

- Some modern sites that allow this and promote it
- Duolingo.com
  - Requires users to do text translation but lets you learn a language



2. Word Classes

- Basic Approach
  - R. Lins & P. Goncalves (2004) ACM Symposium on Applied Computing
- Word Classes
  - Adverbs
  - Articles
  - Conjunctions
  - Interjections
  - Numerals
  - Prepositions
  - Pronouns



Algorithm – 1 Word Class

- Requires dictionary of detected language
  - English, French, Port, Spanish
- Compute
  - Importance/ratio of # of eng, french, spanish port words to overall total (use 1 class)
  - Freq of input text words from class to those in the specific lang dict
  - Check in this order: English, then Port, then French, else Spanish

```
01 idiom(english,french,portuguese,spanish,qtWords)
02 total=english+portuguese+spanish+french
03 if(total==0)return "NEXIST"
04 engId=(english/total)*100
05 porId=(portuguese/total)*100
06 freId=(french/total)*100
07 spaId=(spanish/total)*100
08 engPal=(english/qtWords)*100
09 porPal=(portuguese/qtWords)*100
10 frePal=(french/qtWords)*100
11 spaPal=(spanish/qtWords)*100
12 LimID=40
13 LimPAL=5
14 if(english>(portuguese,spanish,french) &
     engId>=LimID & engPal>=LimPAL) return "ENGLISH"
16 else if(portuguese>(spanish,french) & porId>=LimID
     & porPal>=LimPAL) return "PORTUGUESE"
18 else if(french>spanish & freId>=LimID &
     frePal>=LimPAL)return "FRENCH"
20 else if(spaId>=LimID & spaPal>=LimPAL)
21 return "SPANISH"
22 else return "UNKNOWN"
```

Algorithm – 1 Word Class

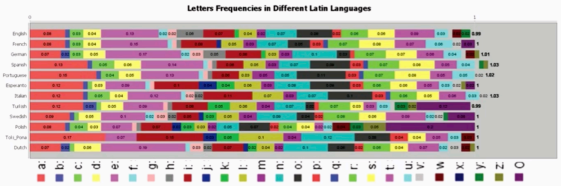
- If it's not English, try Portuguese, otherwise French, otherwise Spanish
- Developed priority order from examining web documents and tuning
- This was repeated for all the Word Classes mentioned

Word Class – 1 class - Results

- Prepositions – mostly classified as Spanish (72%), did well on French, did well on English (72% - 92% overall accuracy)
- Interjections – poorly on French, Portuguese, English
- Number – poorly on French, Portuguese, English
- Adverb – good for French, Portuguese
- Article – good for French, Portuguese, English
- Pronoun – good for French, Portuguese
- Conjunction – poorly for Portuguese
- Updated to use 2 word classes, Preposition, then Article

3. Characters and Statistical Text Properties

- Characters and their frequency is a good signature/indicator of language



N-grams as statistical indicators

- 3-grams of hello
  - hel, ell, llo, _he, lo_ (includes boundary words)
  - In this case, have stored 15 chars instead of 5, so 3x explosion in characters
- Mostly effective on European languages
  - Isolate the important word segments and can be used to compute statistical frequency and as indicators to compare text against
- 3-grams (or more generally N-grams) are compared against input processed text in similar way

Language Identifier Evolution Tika

- Tika 1.x branch
  - Language Identifier was a concrete class and only based on N-Grams
- Overall that approach wasn't that great in practice
  - Computationally efficient and easy, but typically poorly in detecting non US-Latin oriented languages
- Tika 2.x branch
  - Language Identifier being made more modular
  - Integrating Google's Language Identifier and Text.jl as plugins
  - https://issues.apache.org/jira/browse/TIKA-1696
  - https://issues.apache.org/jira/browse/TIKA-1723

Overall concept of "Language Profiling"

- In the case of the word class approach we used classes of "words"
- What's the minimum unit for language identification
- Looking at Dunning 1994
  - Theoretical background for N-gram based statistical identification of text

Unique Letter Combinations

- Chu, 1994
- Enumeration of short sequences that are distinct to language
- Of course, not perfect
  - Zucchini, Pinocchio => not Italian (also English)
  - Amherst, Elmhurst => not only Gaelic
- But decent

N-grams in Tika: As an example

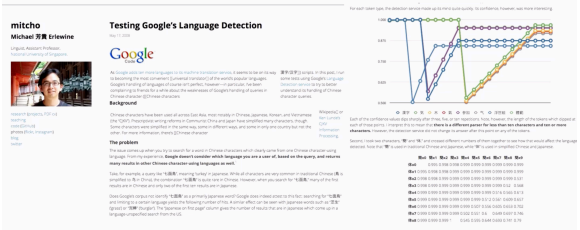
- Given
  - Hello, my name is Chris
  - Text processing:
    - Hello my name is Chris
  - Stopword removal
    - Hello my name Chris
  - N-gram expansion
    - _he, hel, ell, llo, lo_ _my, my_, _na, nam, ame, me_, _Ch, Chr, hri, ris, is_
  - Lowercasing
- Now, take the tokens and compare to language profiles
- Compare
  - Frequency of number of matches (perhaps normalized by max match per profile)
  - Pick the best from there

Some language identification Options

- Optimaize Language Detector
  - https://github.com/optimaize/language-detector
  - N-grams profiles
  - Java
- Compact Language Detector 2
  - https://github.com/CLD2Owners/cld2
  - Written in C
  - Unigrams, with Bayesian probabilistic detection
- MIT Text.jl
  - https://github.com/mit-nlp/Text.jl
  - Language ID – based on N-Grams
  - Written in Julia

Chinese/Eastern Languages

- https://mitcho.com/blog/observation/testing-googles-language-detection/



mitcho
Michael 芳貴 Erlewine

Testing Google's Language Detection

Google

Background

The problem

Summary

- Why Language Identification / Detection
- Representing and Standardized Language Codes (ISO-619)
- Language Identification Approaches
  - By Tagging / Humans
  - Automatic
    - Word Class
    - N-grams
- Libraries for Language Detection

NIST Language Detection Evaluation

- National Institutes of Standards and Technology (NIST)
- Runs Language Detection Evaluation – last was in 2007

ISCA Archive
http://www.isca-speech.org/archive

Odyssey 2008: The Speaker and Language Recognition Workshop
Stellenbosch, South Africa
January 21-24, 2008

NIST 2007 Language Recognition Evaluation

Alvin F. Martin, Audrey N. Le

Speech Group, Information Access Division, Information Technology Laboratory
National Institute of Standards and Technology, USA
{alvin.martin,audrey.le}@nist.gov

Key Resources

- Tika in Action, Chapter 7
- R. Lins & P. Gonclaves, 2004 ACM Symposium on Applied Computing
- T. Dunning. Statistical Identification in Language
- B. Martis and M. J. Silva. Language Identification in Web Pages