Introducing Mallet

http://mallet.cs.umass.edu/mallet-tutorial.pdf
http://programminghistorian.org/lessons/topic-modeling-and-mallet

Mallet

- Java code

- text data: ---> discrete values!

  document classification
  clustering
  information extraction
  ...and other ML applications to text

Models for Text Data

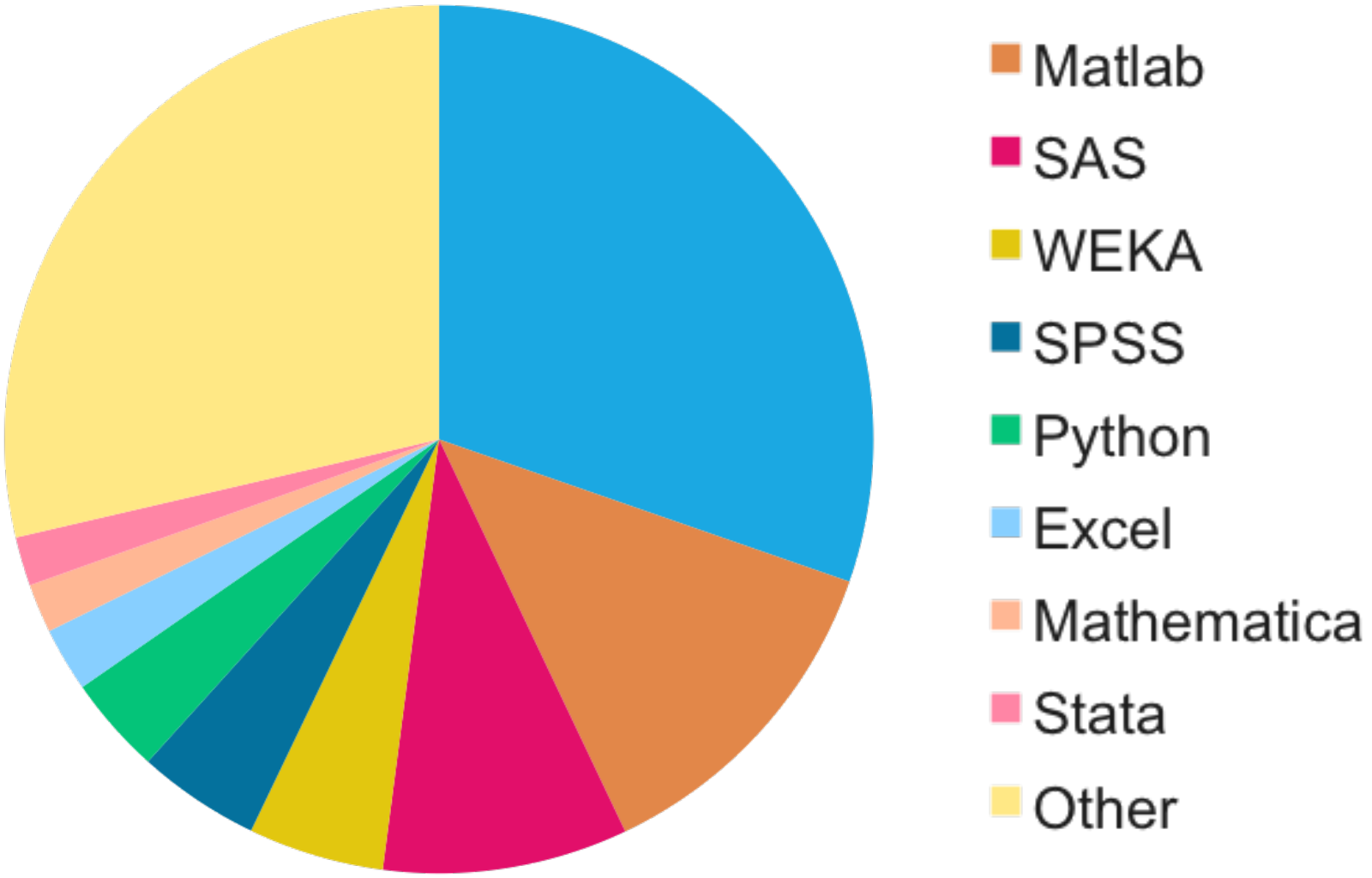
- generative models
  - naive bayes
  - Hidden Markov Models (HMMs)
  - Latent Dirichlet Topic Models

- discriminative regression models
  - MaxEnt/Logistic regression
  - Conditional Random Fields (CRFs)

Features

- Classification
- Sequence Tagging
- Topic Modeling
- Multi-language Support

Why Mallet


R
Matlab
SAS
WEKA
SPSS
Python
Excel
Mathematica
Stata
Other

Why Mallet

- research:
  - robust and fast implementations
  - popular among comp. linguists and digital humanities crowd

- deployment
  - “if you're planning on integrating mallet into another project you should be very familiar with Java and ready to spend a lot of time debugging an almost completely undocumented code base”

How?

- command line scripts
  bin/mallet [command] --[option] [value]

- Direct Java API
  http://mallet.cs.umass.edu/api/

Downloading & installing

- http://mallet.cs.umass.edu/

- prerequisite: ant

Example: topic modeling

- command line

- Java code

http://mallet.cs.umass.edu/topics-devel.php

The good, the bad and the ugly

- one of the leading academic tools for text classification, topic modeling, and sequential tagging using CRF

- no GUI

- NLTK borrows from Mallet