Machine Learning with MALLET

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

David Mimno

Department of Information Science,
Cornell University

Outline

- About MALLET
- Representing Data
- Classification
- Sequence Tagging
- Topic Modeling

Outline

- About MALLET
- Representing Data
- Classification
- Sequence Tagging
- Topic Modeling

Who?

- Andrew McCallum (most of the work)
- Charles Sutton, Aron Culotta, Greg Druck, Kedar Bellare, Gaurav Chandalia…
- Fernando Pereira, others at Penn…

Who am I?

- Chief maintainer of MALLET
- Primary author of MALLET topic modeling package

Why?

- Motivation: text classification and information extraction
- Commercial machine learning (Just Research, WhizBang)
- Analysis and indexing of academic publications: Cora, Rexa

What?

- Text focus: data is discrete rather than continuous, even when values *could* be continuous:

```
double value = 3.0
```

How?

- Command line scripts:
  - bin/mallet [command] --[option] [value] …
  - Text User Interface ("tui") classes
- Direct Java API
  - http://mallet.cs.umass.edu/api

Most of this talk

History

- Version 0.4: c2004
  - Classes in edu.umass.cs.mallet.base.*
- Version 2.0: c2008
  - Classes in cc.mallet.*
  - Major changes to finite state transducer package
  - bin/mallet vs. specialized scripts
  - Java 1.5 generics

Learning More

- http://mallet.cs.umass.edu
  - "Quick Start" guides, focused on command line processing
  - Developers' guides, with Java examples
- mallet-dev@cs.umass.edu mailing list
  - Low volume, but can be bursty

Outline

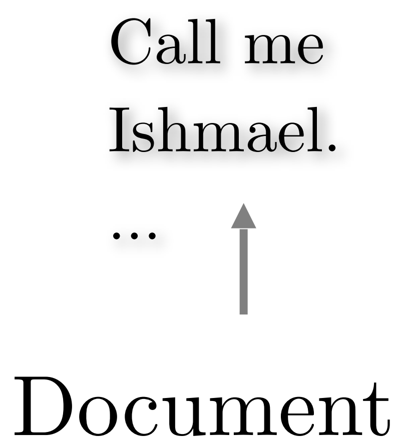
- About MALLET
- Representing Data
- Classification
- Sequence Tagging
- Topic Modeling

Models for Text Data

- Generative models (Multinomials)
  - Naïve Bayes
  - Hidden Markov Models (HMMs)
  - Latent Dirichlet Topic Models
- Discriminative Regression Models
  - MaxEnt/Logistic regression
  - Conditional Random Fields (CRFs)

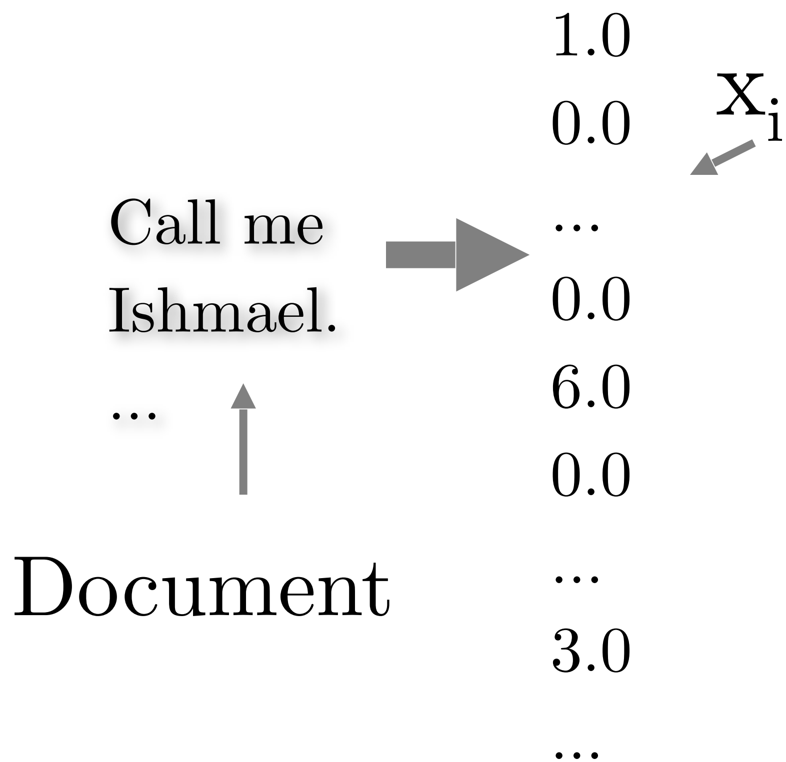
Representations

- Transform text documents to vectors $\mathbf{x}_1$, $\mathbf{x}_2$,…
- Retain meaning of vector indices
- Ideally sparsely



Representations

- Transform text documents to vectors $\mathbf{x_1}$, $\mathbf{x_2}$,…
- Retain meaning of vector indices
- Ideally sparsely



Representations

- Elements of vector are called **feature values**
- Example: Feature at row 345 is number of times "dog" appears in document

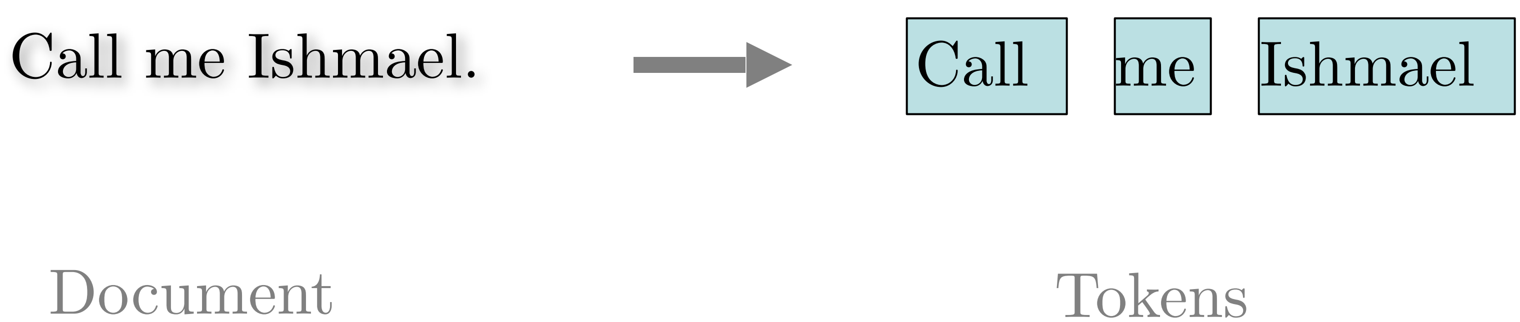
$$\begin{array}{l} 1.0 \\ 0.0 \\ \ldots \\ 0.0 \\ 6.0 \\ 0.0 \\ \ldots \\ 3.0 \\ \ldots \end{array} \qquad \mathrm{x_i}$$

Documents to Vectors

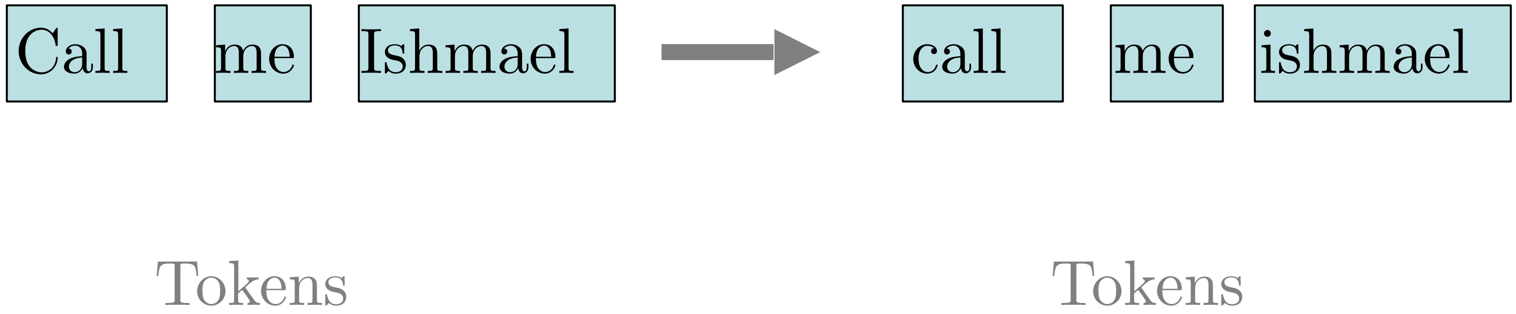
Call me Ishmael.

Document

Documents to Vectors



Documents to Vectors



Documents to Vectors

call me ishmael → 473, 3591, 17

Tokens

Features

17.	ishmael
...	
473.	call
...	
3591	me

Documents to Vectors

473, 3591, 17 →

17	1.0
473	1.0
3591	1.0

Features (sequence)

Features (bag)

17.	ishmael
...	
473.	call
...	
3591	me

17.	ishmael
...	
473.	call
...	
3591	me

Instances

Email message, web page, sentence, journal abstract…

What is it called?

- Name

What is the input?

- Data
- Target/Label

What is the output?

- Source

What did it originally look like?

Instances

- Name
- Data →
- Target
- Source

```
String

TokenSequence
    ArrayList<Token>
FeatureSequence
    int[]
FeatureVector
    int -> double map
```

cc.mallet.types

Alphabets

17.	ishmael
...	
473.	call
...	
3591	me

```
TObjectIntHashMap map
ArrayList entries
```

```
int lookupIndex(Object o, boolean shouldAdd)
```

```
Object lookupObject(int index)
```

cc.mallet.types, gnu.trove

Alphabets

17.	ishmael
...	
473.	call
...	
3591	me

for

```
TObjectIntHashMap map
ArrayList entries
```

```
int lookupIndex(Object o, boolean shouldAdd)

Object lookupObject(int index)
```

cc.mallet.types, gnu.trove

Alphabets

17.	ishmael
...	
473.	call
...	
3591	me

```
TObjectIntHashMap map
ArrayList entries
```

```
void stopGrowth()
```

Do not add entries for new Objects -- default is to allow growth.

```
void startGrowth()
```

cc.mallet.types, gnu.trove

Creating Instances

- Instance constructor method

```
new Instance(data, target,
                 name, source)
```

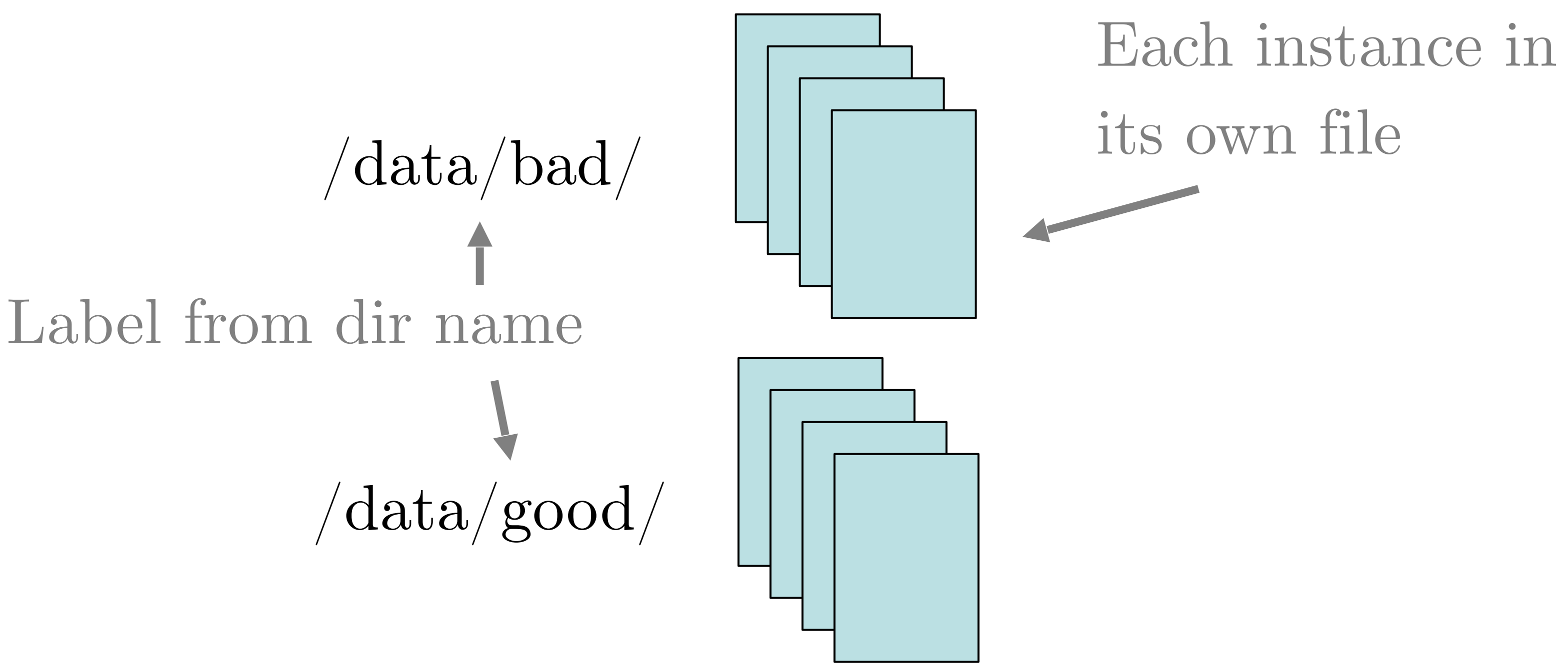
- Iterators

```
Iterator<Instance>
   FileIterator(File[], …)
   CsvIterator(FileReader, Pattern…)
   ArrayIterator(Object[])
   …
```

cc.mallet.pipe.iterator

Creating Instances

- FileIterator



cc.mallet.pipe.iterator

Creating Instances

- CsvIterator

Each instance on its own line

1001.	Melville	Call me Ishmael. Some years ago...
1002.	Dickens	It was the best of times, it was...

^([^\t]+)\t([^\t]+)\t(.*)

Name, label, data from regular expression groups.
"CSV" is a lousy name. LineRegexIterator?

cc.mallet.pipe.iterator

Instance Pipelines

- Sequential transformations of instance fields (usually Data)
- Pass an ArrayList<Pipe> to SerialPipes

```
// “data” is a String
CharSequence2TokenSequence
// tokenize with regexp
TokenSequenceLowercase
// modify each token’s text
TokenSequenceRemoveStopwords
// drop some tokens
TokenSequence2FeatureSequence
// convert token Strings to ints
FeatureSequence2FeatureVector
// lose order, count duplicates
```

cc.mallet.pipe

Instance Pipelines

- A small number of pipes modify the “target” field
- There are now two alphabets: data and label

```
// “target” is a String
Target2Label
// convert String to int
// “target” is now a Label
```

```
Alphabet > LabelAlphabet
```

cc.mallet.pipe, cc.mallet.types

Label objects

- Weights on a fixed set of classes
- For training data, weight for correct label is 1.0, all others 0.0

```
implements Labeling

int getBestIndex()
Label getBestLabel()
```

You cannot create a `Label`, they are only produced by `LabelAlphabet`

cc.mallet.types

InstanceLists

- A List of Instance objects, along with a Pipe, data Alphabet, and LabelAlphabet

```
InstanceList instances =
    new InstanceList(pipe);

instances.addThruPipe(iterator);
```

cc.mallet.types

Putting it all together

```
ArrayList<Pipe> pipeList = new ArrayList<Pipe>();

pipeList.add(new Target2Label());
pipeList.add(new CharSequence2TokenSequence());
pipeList.add(new TokenSequence2FeatureSequence());
pipeList.add(new FeatureSequence2FeatureVector());

InstanceList instances =
   new InstanceList(new SerialPipes(pipeList));

instances.addThruPipe(new FileIterator(. . .));
```

Persistent Storage

- Most MALLET classes use Java serialization to store models and data

```
ObjectOutputStream oos =
    new ObjectOutputStream(…);
oos.writeObject(instances);
oos.close();
```

Pipes, data objects, labelings, etc all need to implement Serializable.

Be sure to include custom classes in classpath, or you get a StreamCorruptedException

java.io

Review

- What are the four main fields in an Instance?

Review

- What are the four main fields in an Instance?
- What are two ways to generate Instances?

Review

- What are the four main fields in an Instance?
- What are two ways to generate Instances?
- How do we modify the value of Instance fields?

Review

- What are the four main fields in an Instance?
- What are two ways to generate Instances?
- How do we modify the value of Instance fields?
- Name some classes that appear in the "data" field.

Outline

- About MALLET
- Representing Data
- Classification
- Sequence Tagging
- Topic Modeling

Classifier objects

- Classifiers map from instances to distributions over a fixed set of classes
- MaxEnt, Naïve Bayes, Decision Trees…

Given data

watery

NN
JJ
PRP
VB
CC

Which class is best?

(this one!)

cc.mallet.classify

Classifier objects

- Classifiers map from instances to distributions over a fixed set of classes
- MaxEnt, Naïve Bayes, Decision Trees…

```
Labeling labeling =
    classifier.classify(instance);

Label l = labeling.getBestLabel();

System.out.print(instance + “\t”);
System.out.println(l);
```

cc.mallet.classify

Training Classifier objects

- Each type of classifier has one or more ClassifierTrainer classes

```
ClassifierTrainer trainer =
    new MaxEntTrainer();

Classifier classifier =
    trainer.train(instances);
```

Training Classifier objects

- Some classifiers require numerical optimization of an objective function.

$$
\begin{aligned}
&\log \mathrm{P}(\text{Labels} \mid \text{Data}) = \\
&\quad \log \mathrm{f}(\text{label}_1, \text{data}_1, \mathrm{w}) + \\
&\quad \log \mathrm{f}(\text{label}_2, \text{data}_2, \mathrm{w}) + \\
&\quad \log \mathrm{f}(\text{label}_3, \text{data}_3, \mathrm{w}) + \\
&\quad\ \ldots
\end{aligned}
$$

Maximize w.r.t. w!

cc.mallet.optimize

Parameters **w**

- Association between feature, class label
- How many parameters for K classes and N features?

action	NN	0.13
action	VB	-0.1
action	JJ	-0.21
SUFF-tion	NN	1.3
SUFF-tion	VB	-2.1
SUFF-tion	JJ	-1.7
SUFF-on	NN	0.01
SUFF-on	VB	-0.02
…		

Training Classifier objects

```
interface Optimizer
   boolean optimize()

interface Optimizable
   interface ByValue
   interface ByValueGradient
```

Limited-memory BFGS,
Conjugate gradient...

Specific objective functions

cc.mallet.optimize

Training Classifier objects

```
MaxEntOptimizableByLabelLikelihood
    double[] getParameters()
    void setParameters(double[] parameters)
    …

    double getValue()
    void getValueGradient(double[] buffer)
```

For Optimizable interface

Log likelihood and its first derivative

cc.mallet.classify

Evaluation of Classifiers

- Create random test/ train splits

```
InstanceList[] instanceLists =
   instances.split(new Randoms(),
             new double[] {0.9, 0.1, 0.0});
```

90% training

10% testing

0% validation

cc.mallet.types

Evaluation of Classifiers

- The Trial class stores the results of classifications on an InstanceList (testing or training)

```
Trial(Classifier c, InstanceList list)
    double getAccuracy()
    double getAverageRank()
    double getF1(int/Label/Object)
    double getPrecision(…)
    double getRecall(…)
```

cc.mallet.classify

Review

- I have invented a new classifier: David regression.
  - What class should I implement to classify instances?

Review

- I have invented a new classifier: David regression.
  - What class should I implement to train a David regression classifier?

Review

- I have invented a new classifier: David regression.
  - I want to train using ByValueGradient. What mathematical functions do I need to code up, and what class should I put them in?

Review

- I have invented a new classifier: David regression.
  - How would I check whether my new classifier works better than Naïve Bayes?

Outline

- About MALLET
- Representing Data
- Classification
- Sequence Tagging
- Topic Modeling

Sequence Tagging

- Data occurs in sequences
- Categorical labels for each position
- Labels are correlated

DET	NN	VBS	VBG
the	dog	likes	running

Sequence Tagging

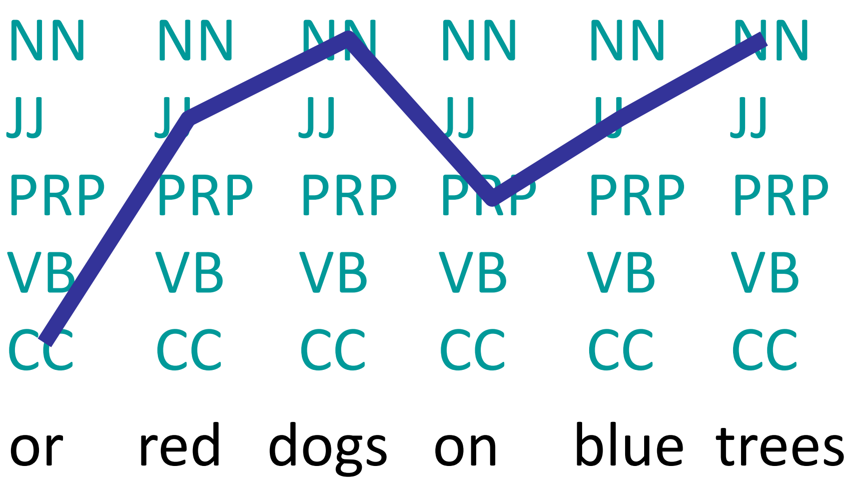
- Data occurs in sequences
- Categorical labels for each position
- Labels are correlated

??	??	??	??
the	dog	likes	running

Sequence Tagging

- Classification: n-way
- Sequence Tagging: n^T-way

NN
JJ
PRP
VB
CC



Avoiding Exponential Blowup

- Markov property
- Dynamic programming

Andrei Markov

Avoiding Exponential Blowup

- Markov property
- Dynamic programming

DET JJ NN VB

This one

Given this one

Is independent of these

Andrei Markov

Avoiding Exponential Blowup

- Markov property
- Dynamic programming

NN	NN	NN	NN	NN	NN
JJ	JJ	JJ	JJ	JJ	JJ
PRP	PRP	PRP	PRP	PRP	PRP
VB	VB	VB	VB	VB	VB
CC	CC	CC	CC	CC	CC
or	red	dogs	on	blue	trees

Andrei Markov

Avoiding Exponential Blowup

- Markov property
- Dynamic programming

NN	NN	NN	NN	NN
JJ	JJ	JJ	JJ	JJ
PRP	PRP	PRP	PRP	PRP
VB	VB	VB	VB	VB
CC	CC	CC	CC	CC
red	dogs	on	blue	trees

Andrei Markov

Avoiding Exponential Blowup

- Markov property
- Dynamic programming

NN	NN	NN	NN
JJ	JJ	JJ	JJ
PRP	PRP	PRP	PRP
VB	VB	VB	VB
CC	CC	CC	CC
dogs	on	blue	trees

Andrei Markov

Hidden Markov Models and Conditional Random Fields

- Hidden Markov Model: fully generative

$$\mathrm{P(Labels \mid Data) =}$$
$$\mathrm{P(Data, Labels)\ /\ P(Data)}$$

- Conditional Random Field: conditional

$$\mathrm{P(Labels \mid Data)}$$

Hidden Markov Models and Conditional Random Fields

- Hidden Markov Model: simple (independent) output space

“NSF-funded”

- Conditional Random Field: arbitrarily complicated outputs

“NSF-funded”
CAPITALIZED
HYPHENATED
ENDS-WITH-ed
ENDS-WITH-d
…

Hidden Markov Models and Conditional Random Fields

- Hidden Markov Model: simple (independent) output space

`FeatureSequence`

`int[]`

- Conditional Random Field: arbitrarily complicated outputs

`FeatureVectorSequence`

`FeatureVector[]`

Importing Data

- SimpleTagger format: one word per line, with instances delimited by a blank line

Call VB
me PPN
Ishmael NNP
. .

Some JJ
years NNS
...

Importing Data

- SimpleTagger format: one word per line, with instances delimited by a blank line

```
Call SUFF-ll VB
me TWO_LETTERS PPN
Ishmael BIBLICAL_NAME NNP
. PUNCTUATION .

Some CAPITALIZED JJ
years TIME SUFF-s NNS
...
```

Importing Data

```
LineGroupIterator

    SimpleTaggerSentence2TokenSequence()
    //String to Tokens, handles labels

    TokenSequence2FeatureVectorSequence()
    //Token objects to FeatureVectors
```

cc.mallet.pipe, cc.mallet.pipe.iterator

Importing Data

```
LineGroupIterator

   SimpleTaggerSentence2TokenSequence()
   //String to Tokens, handles labels

     [Pipes that modify tokens]

   TokenSequence2FeatureVectorSequence()
   //Token objects to FeatureVectors
```

cc.mallet.pipe, cc.mallet.pipe.iterator

Importing Data

must match
entire string

```
   //Ishmael
TokenTextCharSuffix("C2=", 2)
   //Ishmael C2=el
RegexMatches("CAP", Pattern.compile("\\p{Lu}.*"))
   //Ishmael C2=el CAP
LexiconMembership("NAME", new File('names'), false)
   //Ishmael C2=el CAP NAME
```

one name per line

ignore case?

cc.mallet.pipe.tsf

Sliding window features

a red $\boxed{\text{dog}}$ on a blue tree

Sliding window features

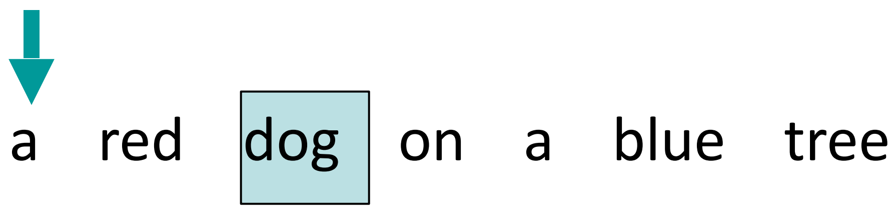
a red dog on a blue tree

Sliding window features

a red $\boxed{\text{dog}}$ on a blue tree

red@-1

Sliding window features



red@-1

a@-2

Sliding window features

a red dog on a blue tree

red@-1

a@-2

on@1

Sliding window features

a red dog on a blue tree

red@-1

a@-2

on@1

a@-2_&_red@-1

Importing Data

```
int[][] conjunctions = new int[3][];
      conjunctions[0] = new int[] { -1 };
      conjunctions[1] = new int[] { 1 };
      conjunctions[2] = new int[] { -2, -1 };

OffsetConjunctions(conjunctions)

   // a@-2_&_red@-1 on@1
```

previous position

next position

previous two

cc.mallet.pipe.tsf

Importing Data

previous position

next position

previous two

```
int[][] conjunctions = new int[3][];
        conjunctions[0] = new int[] { -1 };
        conjunctions[1] = new int[] { 1 };
        conjunctions[2] = new int[] { -2, -1 };

TokenTextCharSuffix("C1=", 1)
OffsetConjunctions(conjunctions)

   // a@-2_&_red@-1 a@-2_&_C1=d@-1
```

cc.mallet.pipe.tsf

Finite State Transducers

- Finite state machine over two alphabets (observed, hidden)

Finite State Transducers

- Finite state machine over two alphabets (observed, hidden)

DET

$\mathrm{P}(\mathrm{DET})$

Finite State Transducers

- Finite state machine over two alphabets (observed, hidden)

DET
the

$$\mathrm{P}(\mathrm{the} \mid \mathrm{DET})$$

Finite State Transducers

- Finite state machine over two alphabets (observed, hidden)

DET NN
the

$$\mathrm{P}(\mathrm{NN} \mid \mathrm{DET})$$

Finite State Transducers

- Finite state machine over two alphabets (observed, hidden)

DET NN
the dog

$$\mathrm{P}(\mathrm{dog} \mid \mathrm{NN})$$

Finite State Transducers

- Finite state machine over two alphabets (observed, hidden)

DET NN VBS
the dog

$$\mathrm{P}(\mathrm{VBS} \mid \mathrm{NN})$$

How many parameters?

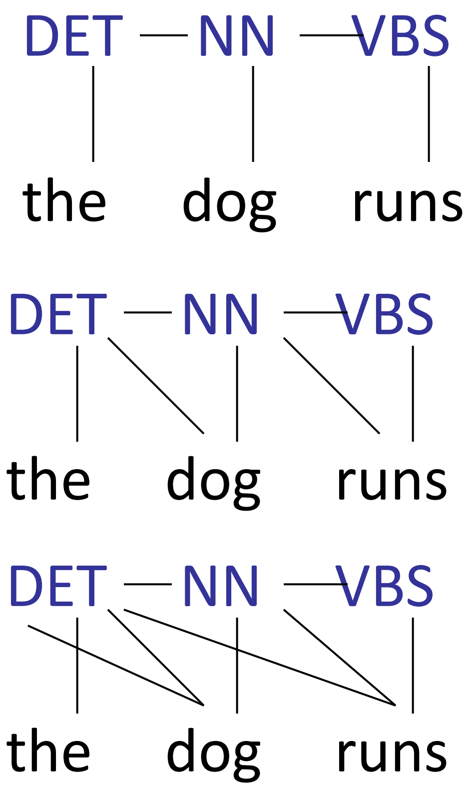
- Determines efficiency of training
- Too many leads to overfitting

Trick: Don't allow certain transitions

$$P(\text{VBS} \mid \text{DET}) = 0$$

How many parameters?

- Determines efficiency of training
- Too many leads to overfitting



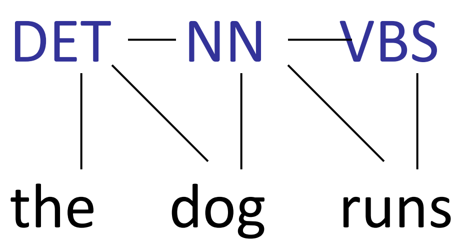
Finite State Transducers

```
abstract class Transducer
   CRF
   HMM

abstract class TransducerTrainer
   CRFTrainerByLabelLikelihood
   HMMTrainerByLikelihood
```

cc.mallet.fst

Finite State Transducers

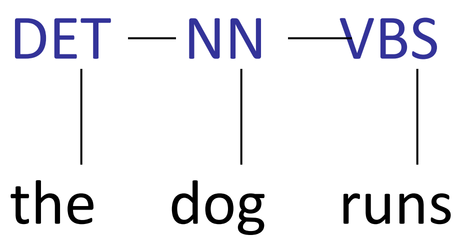


First order: one weight for every pair of labels and observations.

```
CRF crf = new CRF(pipe, null);
crf.addFullyConnectedStates();
     // or
crf.addStatesForLabelsConnectedAsIn(instances);
```

cc.mallet.fst

Finite State Transducers

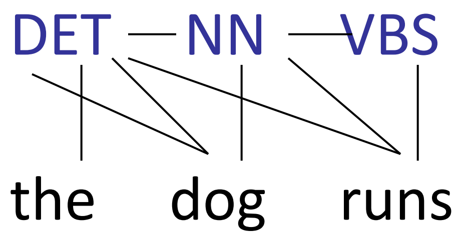


"three-quarter" order: one weight for every pair of labels and observations.

```
crf.addStatesForThreeQuarterLabelsConnectedAsIn(instances);
```

cc.mallet.fst

Finite State Transducers



Second order: one weight for every triplet of labels and observations.

```
crf.addStatesForBiLabelsConnectedAsIn(instances);
```

cc.mallet.fst

Finite State Transducers

DET	NN	VBS
the	dog	runs

"Half" order: equivalent to independent classifiers, except some transitions may be illegal.

```
crf.addStatesForHalfLabelsConnectedAsIn(instances);
```

cc.mallet.fst

Training a transducer

```
CRF crf = new CRF(pipe, null);
crf.addStatesForLabelsConnectedAsIn(trainingInstances);

CRFTrainerByLabelLikelihood trainer =
    new CRFTrainerByLabelLikelihood(crf);

trainer.train();
```

cc.mallet.fst

Evaluating a transducer

```
CRFTrainerByLabelLikelihood trainer =
    new CRFTrainerByLabelLikelihood(transducer);

TransducerEvaluator evaluator =
    new TokenAccuracyEvaluator(testing, "testing"));

trainer.addEvaluator(evaluator);

trainer.train();
```

Applying a transducer

```
Sequence output = transducer.transduce (input);

for (int index=0; index < input.size(); input++) {
      System.out.print(input.get(index) + "/");
      System.out.print(output.get(index) + " ");
}
```

cc.mallet.fst

Review

- How do you add new features to TokenSequences?

Review

- How do you add new features to TokenSequences?
- What are three factors that affect the number of parameters in a model?

Outline

- About MALLET
- Representing Data
- Classification
- Sequence Tagging
- Topic Modeling

Topics: “Semantic Groups”

News Article

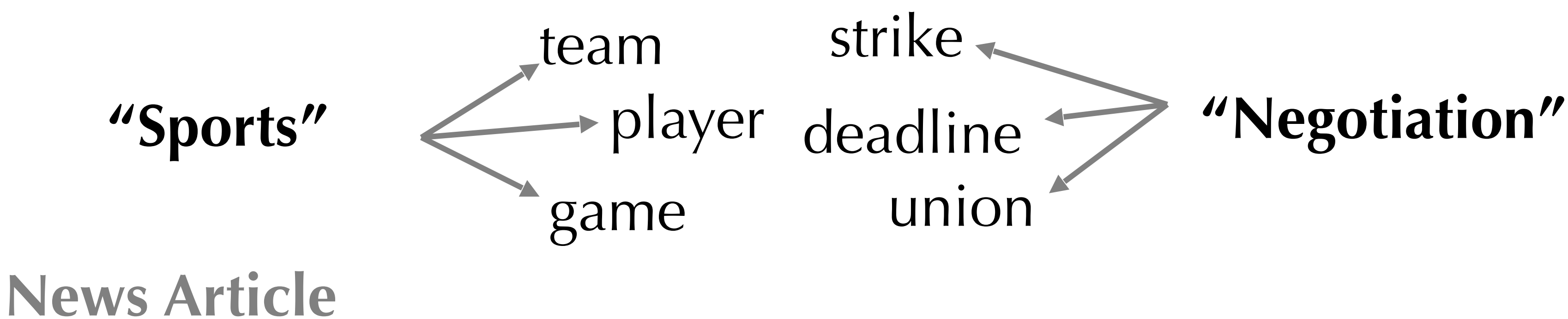
Topics: “Semantic Groups”

“Sports”

“Negotiation”

News Article

Topics: “Semantic Groups”



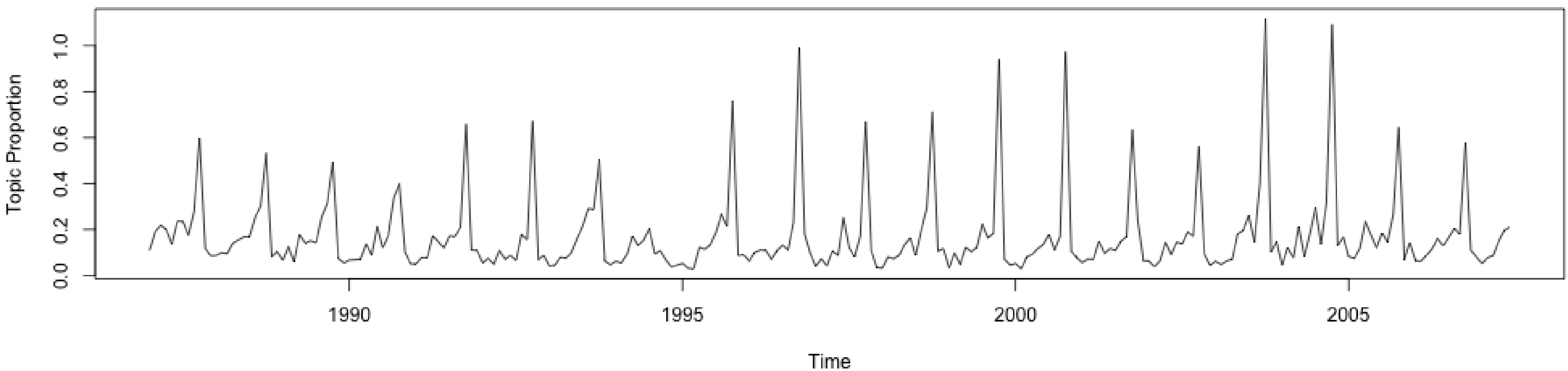
Topics: “Semantic Groups”

team
player
game

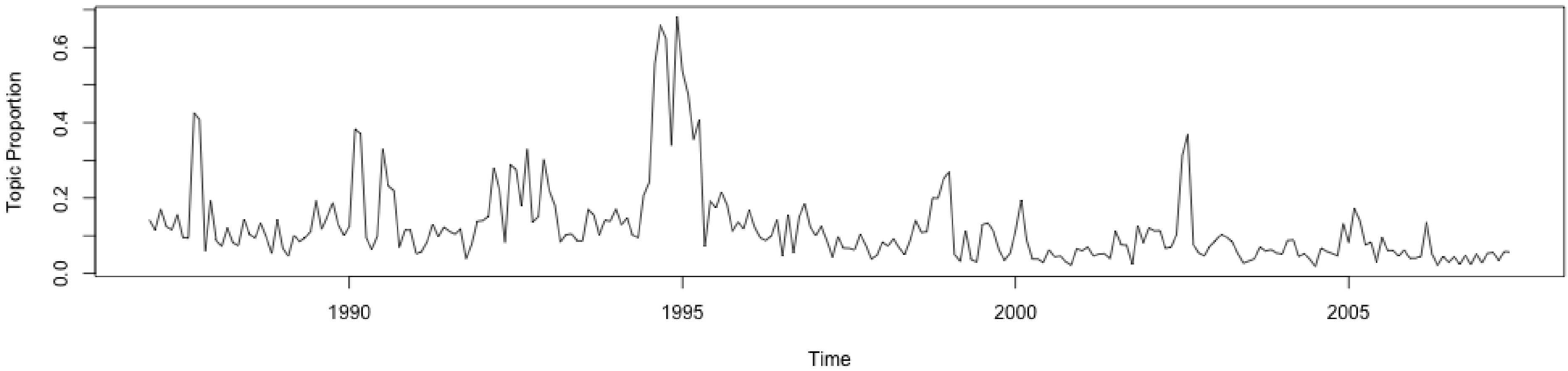
strike
deadline
union

News Article

Series Yankees Sox Red World League game Boston team games baseball Mets Game series won Clemens Braves Yankee teams



players League owners league baseball union commissioner Baseball Association labor Commissioner Football major teams Selig agreement strike team bargaining



Training a Topic Model

```
ParallelTopicModel lda = new ParallelTopicModel(numTopics);
lda.addInstances(trainingInstances);
lda.estimate();
```

cc.mallet.topics

Evaluating a Topic Model

```
ParallelTopicModel lda = new ParallelTopicModel(numTopics);
lda.addInstances(trainingInstances);
lda.estimate();

MarginalProbEstimator evaluator =
   lda.getProbEstimator();

double logLikelihood =
   evaluator.evaluateLeftToRight(testing, 10, false, null);
```

cc.mallet.topics

Inferring topics for new documents

```
ParallelTopicModel lda = new ParallelTopicModel(numTopics);
lda.addInstances(trainingInstances);
lda.estimate();

TopicInferencer inferencer =
    lda.getInferencer();

double[] topicProbs =
    inferencer.getSampledDistribution(instance, 100,
                                      10, 10);
```

cc.mallet.topics

More than words…

- Text collections mix free text and structured data

David Mimno
Andrew McCallum
UAI
2008
…

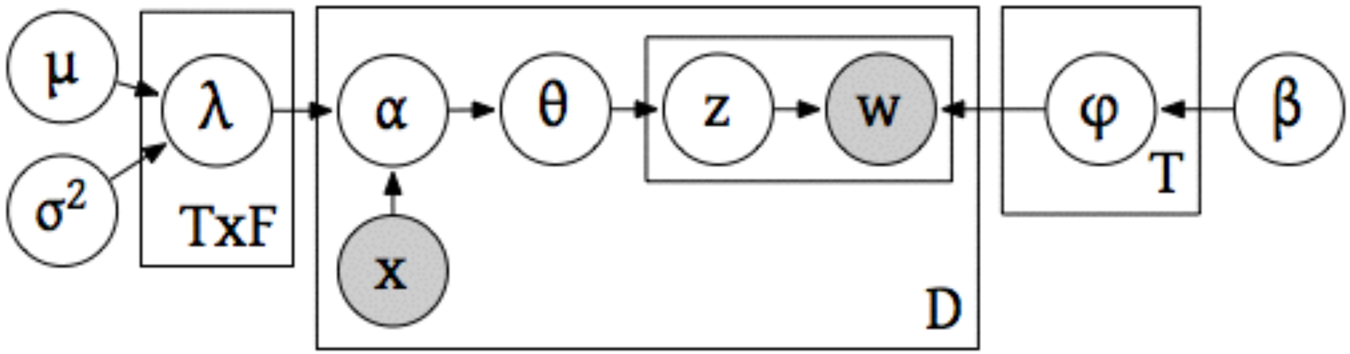
More than words…

- Text collections mix free text and structured data

David Mimno
Andrew McCallum
UAI
2008

"Topic models conditioned on arbitrary features using Dirichlet-multinomial regression. …"

Dirichlet-multinomial Regression (DMR)



The corpus specifies a vector of real-valued features (x) for each document, of length F.

Each topic has an F-length vector of parameters.

Topic parameters for feature “published in JMLR”

2.27	kernel, kernels, rational kernels, string kernels, fisher kernel
1.74	bounds, vc dimension, bound, upper bound, lower bounds
1.41	reinforcement learning, learning, reinforcement
1.40	blind source separation, source separation, separation, channel
1.37	nearest neighbor, boosting, nearest neighbors, adaboost
-1.12	agent, agents, multi agent, autonomous agents
-1.21	strategies, strategy, adaptation, adaptive, driven
-1.23	retrieval, information retrieval, query, query expansion
-1.36	web, web pages, web page, world wide web, web sites
-1.44	user, users, user interface, interactive, interface

Feature parameters for RL topic

2.99	Sridhar Mahadevan
2.88	ICML
2.56	Kenji Doya
2.45	ECML
2.19	Machine Learning Journal
-1.38	ACL
-1.47	CVPR
-1.54	IEEE Trans. PAMI
-1.64	COLING
-3.76	<default>

Topic parameters for feature “published in UAI”

2.88	bayesian networks, bayesian network, belief networks
2.26	qualitative, reasoning, qualitative reasoning, qualitative simulation
2.25	probability, probabilities, probability distributions,
2.25	uncertainty, symbolic, sketch, primal sketch, uncertain, connectionist
2.11	reasoning, logic, default reasoning, nonmonotonic reasoning
-1.29	shape, deformable, shapes, contour, active contour
-1.36	digital libraries, digital library, digital, library
-1.37	workshop report, invited talk, international conference, report
-1.50	descriptions, description, top, bottom, top bottom
-1.50	nearest neighbor, boosting, nearest neighbors, adaboost

Feature parameters for Bayes nets topic

2.88	UAI
2.41	Mary-Anne Williams
2.23	Ashraf M. Abdelbar
2.15	Philippe Smets
2.04	Loopy Belief Propagation for Approximate Inference (Murphy, Weiss, and Jordan, UAI, 1999)
-1.16	Probabilistic Semantics for Nonmonotonic Reasoning (Pearl, KR, 1989)
-1.38	COLING
-1.50	Neural Networks
-2.24	ICRA
-3.36	<default>

Dirichlet-multinomial Regression

- Arbitrary observed features of documents
- Target contains FeatureVector

```
DMRTopicModel dmr =
   new DMRTopicModel (numTopics);

dmr.addInstances(training);
dmr.estimate();

dmr.writeParameters(new File("dmr.parameters"));
```

Polylingual Topic Modeling

- Topics exist in more languages than you could possibly learn
- Topically *comparable* documents are much easier to get than translation sets
- Translation dictionaries
  - cover pairs, not sets of languages
  - miss technical vocabulary
  - aren’t available for low-resource languages

Topics from European Parliament Proceedings

DA	centralbank europæiske ecb s lån centralbanks
DE	zentralbank ezb bank europäischen investitionsbank darlehen
EL	τράπεζα τράπεζας κεντρική εκτ κεντρικής τράπεζες
EN	**bank central ecb banks european monetary**
ES	banco central europeo bce bancos centrales
FI	keskuspankin ekp n euroopan keskuspankki eip
FR	banque centrale bce européenne banques monétaire
IT	banca centrale bce europea banche prestiti
NL	bank centrale ecb europese banken leningen
PT	banco central europeu bce bancos empréstimos
SV	centralbanken europeiska ecb centralbankens s lån

DA	børn familie udnyttelse børns børnene seksuel
DE	kinder kindern familie ausbeutung familien eltern
EL	παιδιά παιδιών οικογένεια οικογένειας γονείς παιδικής
EN	**children family child sexual families exploitation**
ES	niños familia hijos sexual infantil menores
FI	lasten lapsia lapset perheen lapsen lapsiin
FR	enfants famille enfant parents exploitation familles
IT	bambini famiglia figli minori sessuale sfruttamento
NL	kinderen kind gezin seksuele ouders familie
PT	crianças família filhos sexual criança infantil
SV	barn barnen familjen sexuellt familj utnyttjande

Topics from European Parliament Proceedings

DA	mål nå målsætninger målet målsætning opnå
DE	ziel ziele erreichen zielen erreicht zielsetzungen
EL	στόχους στόχο στόχος στόχων στόχοι επίτευξη
EN	**objective objectives achieve aim ambitious set**
ES	objetivo objetivos alcanzar conseguir lograr estos
FI	tavoite tavoitteet tavoitteena tavoitteiden tavoitteita tavoitteen
FR	objectif objectifs atteindre but cet ambitieux
IT	obiettivo obiettivi raggiungere degli scopo quello
NL	doelstellingen doel doelstelling bereiken bereikt doelen
PT	objectivo objectivos alcançar atingir ambicioso conseguir
SV	mål målet uppnå målen målsättningar målsättning

DA	andre anden side ene andet øvrige
DE	anderen andere einen wie andererseits anderer
EL	άλλες άλλα άλλη άλλων άλλους όπως
EN	**other one hand others another there**
ES	otros otras otro otra parte demás
FI	muiden toisaalta muita muut muihin muun
FR	autres autre part côté ailleurs même
IT	altri altre altro altra dall parte
NL	andere anderzijds anderen ander als kant
PT	outros outras outro lado outra noutros
SV	andra sidan å annat ena annan

Topics from Wikipedia

CY	sadwrn blaned gallair at lloeren mytholeg
DE	space nasa sojus flug mission
EL	διαστημικό sts nasa αγγλ small
EN	**space mission launch satellite nasa spacecraft**
FA	فضایی ماموریت ناسا مدار فضانورد ماهواره
FI	sojuz nasa apollo ensimmäinen space lento
FR	spatiale mission orbite mars satellite spatial
HE	החלל הארץ חלל כדור א תוכנית
IT	spaziale missione programma space sojuz stazione
PL	misja kosmicznej stacji misji space nasa
RU	космический союз космического спутник станции
TR	uzay soyuz ay uzaya salyut sovyetler

CY	sbaen madrid el la josé sbaeneg
DE	de spanischer spanischen spanien madrid la
EL	ισπανίας ισπανία de ισπανός ντε μαδρίτη
EN	**de spanish spain la madrid y**
FA	ترین de اسپانیا اسپانیایی کوبا مادرید
FI	espanja de espanjan madrid la real
FR	espagnol espagne madrid espagnole juan y
HE	ספרד ספרדית דה מדריד הספרדית קובה
IT	de spagna spagnolo spagnola madrid el
PL	de hiszpański hiszpanii la juan y
RU	де мадрид испании испания испанский de
TR	ispanya ispanyol madrid la küba real

CY	bardd gerddi iaith beirdd fardd gymraeg
DE	dichter schriftsteller literatur gedichte gedicht werk
EL	ποιητής ποίηση ποιητή έργο ποιητές ποιήματα
EN	**poet poetry literature literary poems poem**
FA	شاعر شعر ادبیات فارسی ادبی آثار
FI	runoilija kirjailija kirjallisuuden kirjoitti runo julkaisi
FR	poète écrivain littérature poésie littéraire ses

Aligned instance lists

dog...	chien...	hund...
cat...	chat...	
pig...		schwein...

Polylingual Topics

```
InstanceList[] training =
    new InstanceList[] { english, german,
                         arabic, mahican };

PolylingualTopicModel pltm =
    new PolylingualTopicModel(numTopics);

pltm.addInstances(training);
```

MALLET hands-on tutorial

http://mallet.cs.umass.edu/mallet-handson.tar.gz