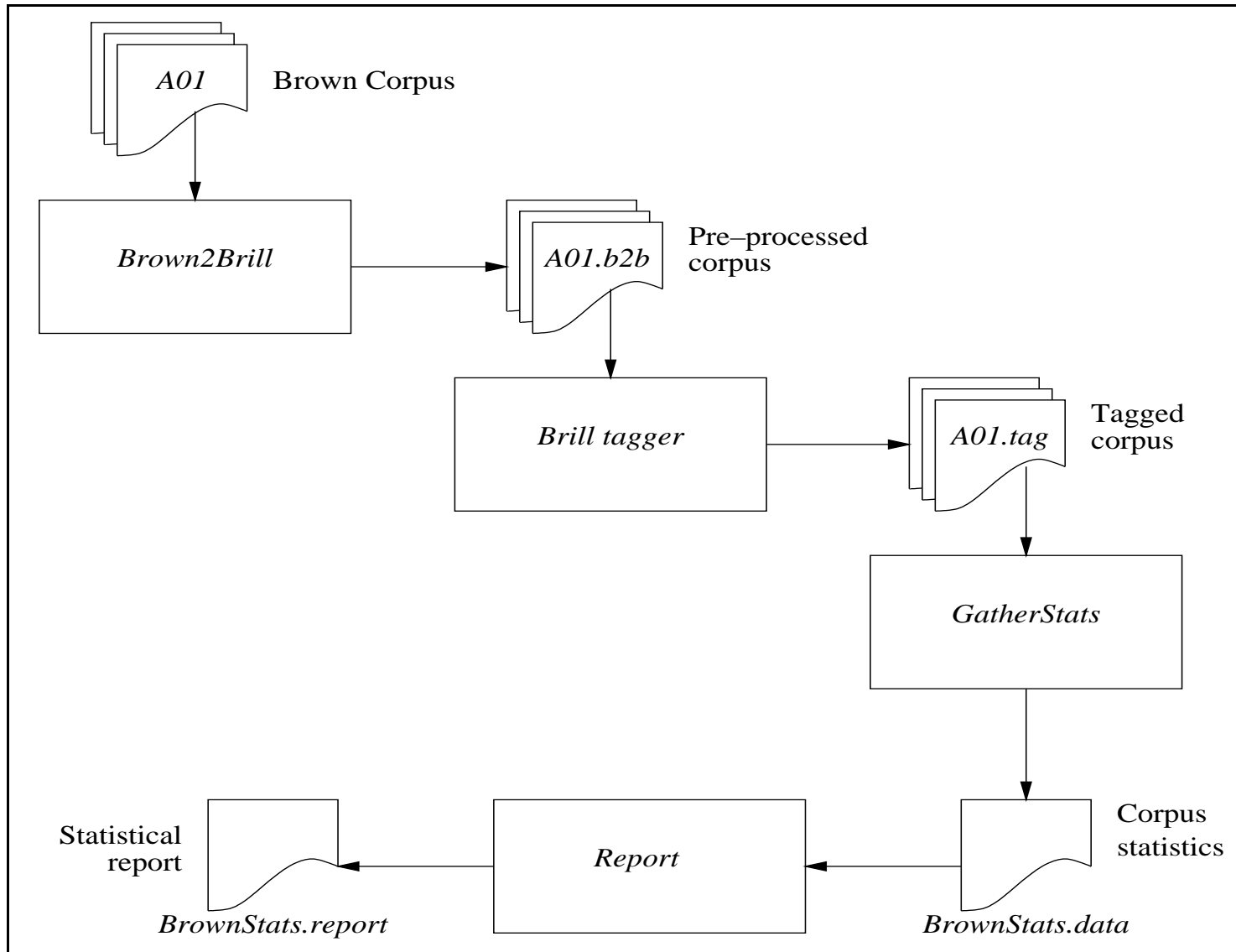




A1 Tutorial

Gathering Statistics from a Corpus



A1 scripts

- Make sure you create the directories you use as arguments.
- How to run the scripts:
 - ▶ `RunBrown2Brill /u/cs401/Brown/corpus corpus.sent`
 - ▶ `RunTagger corpus.sent corpus.tag`
 - ▶ `RunGatherStats corpus.tag BrownStats.data`
- These scripts call your programs:
 - ▶ `Brown2Brill A01 >A01.b2b`
 - ▶ `GatherStats A01.tag >>output_file`

Sentence boundaries

A01 0700 6 The couple was married Aug. 2, 1913. They have a
A01 0710 7 son, William Berry Jr., and a daughter, Mrs. J. M.
A01 0720 4 Cheshire of Griffin.

A01 0720 7 Attorneys for the mayor said that an amicable property
A01 0730 9 settlement has been agreed upon.

A01 0740 2 The petition listed the mayor's occupation as "attorney"
A01 0750 1 and his age as 71. It listed his wife's age as 74 and
A01 0750 14 place of birth as Opelika, Ala.

A01 0760 6 The petition said that the couple has not lived
A01 0770 4 together as man and wife for more than a year.

A01 0780 1 The Hartsfield home is at 637 E. Pelham Rd. NE.

A01 0790 1 Henry L. Bowden was listed on the petition as the
A01 0790 11 mayor's attorney.

Manning & Sckütze's algorithm

- Put potential boundary after all . ? ! ...
- Move the boundary after following " if any
- Disqualify a period boundary if:
 - ▶ preceded by non-sentence final abbreviation, followed by capital letter (Prof., vs.)
 - ▶ preceded by an abbreviation, and not followed by upper case (etc., Jr.)
- Disqualify a ? ! boundary if followed by lower case

Gathering statistics

- E02, 24.50, 12.0, ... , 23.67, 5.84, 0.58, II
- Lists of verbs – need to consider inflected forms:
address/VB, addressing/VBG, addressed/VBD

Type-token ratio

- Number of different types of words in text divided by number of tokens.
- Don't count punctuation tokens.
- Two tokens are the same word-type if they match by case-independent string-equality.
- Same word-types: Horse/NN, horse/NN, horse/VB
- Different word-types: horse/NN, horses/NNS, horsing/VBG

K Nearest Neighbours

- Evaluation mode

For each document D in training data, find K neighbours that are most similar to D , and attribute to D the winning class of the K neighbours. Then see if the computed class for D is its actual class.

- Production mode

Given a new document, label it with the class that is the winning class of its K nearest neighbours.

Cosine measure

Bigger cosine, higher similarity.

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| |\vec{y}|} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

$$\vec{x} = (x_1, x_2, \dots, x_n)$$

$$\vec{y} = (y_1, y_2, \dots, y_n)$$

Precision, recall, and accuracy

(a)	(b)	(c)	(d)	<- classified as
132	6	9	4	(a): class I
2	75	1	6	(b): class II
8	9	119	12	(c): class III
8	3	4	97	(d): class IV

Precision = fraction of cases classified as C that are truly C.

Recall = fraction of cases that truly are C that were classified as C.

Accuracy = fraction of correct decisions.

Table 7 – kNN output

Options:

File stem <BrownStats>

k = 10

10 features used:

First person

Second person

Present-tense verbs

Private verbs

Public verbs

Verbs of saying

Non-existential there

Hifalutin / and technical words

Sentence length

Type-token ratio

Read 495 cases

Table 7 – kNN output

Evaluation on training data:

```
Correct      Errors
-----
      423      72 (14.55%) (accuracy = 85.45%)
```

```
(a)  (b)  (c)  (d)      <- classified as
-----
132   6   9   4      (a): class I
  2  75   1   6      (b): class II
  8   9 119  12     (c): class III
  8   3   4  97     (d): class IV
```

Precision by class:

```
.880 .806 .897 .815 Average = .8495
```

Recall by class:

```
.874 .903 .804 .866 Average = .8618
```

Table 6 – kNN files example

```
| Title: Final settlements in labor negotiations in Canadian industry
| Classes
| -----
good, bad.
| Attributes
| -----
duration:                include
wage increase first year: include
wage increase second year: include
wage increase third year: include
cost of living adjustment: ignore
working hours:           include
pension:                 ignore
standby pay:             include
```

shift differential: include
 education allowance: ignore
 statutory holidays: include
 vacation: ignore
 longterm disability assistance: ignore
 contribution to dental plan: ignore
 bereavement assistance: ignore
 contribution to health plan: ignore

Data:

1,5.0,?,?,?,40,?,?,2,?,11,average,?,?,yes,?,good
 2,4.5,5.8,?,?,35,ret_allw,?,?,yes,11,below average,?,full,?,full,good
 ?,?,?,?,?,38,empl_contr,?,5,?,11,generous,yes,half,yes,half,good
 3,4.5,4.5,5.0,?,40,?,?,?,?,12,average,?,half,yes,half,good
 3,4.0,5.0,5.0,tc,?,empl_contr,?,?,?,12,generous,yes,none,yes,half,good
 3,6.9,4.8,2.3,?,40,?,?,3,?,12,below average,?,?,?,?,good
 2,3.0,7.0,?,?,38,?,12,25,yes,11,below average,yes,half,yes,?,good
 1,5.7,?,?,none,40,empl_contr,?,4,?,11,generous,yes,full,?,?,good