Web 1T 5-gram Corpus Version 1.1
LDC2006T13

1. Introduction

This data set contains English word n-grams and their observed frequency counts. The length of the n-grams ranges from unigrams (single words) to five-grams. We expect this data will be useful for statistical language modeling, e.g., for machine translation or speech recognition, as well as for other uses.

1.1 Source Data

The n-gram counts were generated from approximately 1 trillion word tokens of text from Web pages. We used only publically accessible Web pages. We attempted to use only Web pages with English text, but some text from other languages also found its way into the data.

1.2 Date of Data Collection

Data collection took place in January 2006. This means that no text that was created on or after February 1, 2006 was used.

2. Data Preprocessing

2.1 Character Encoding

The input encoding of documents was automatically detected, and all text was converted to UTF8.

2.2 Tokenization

The data was tokenized in a manner similar to the tokenization of the Wall Street Journal portion of the Penn Treebank. Notable exceptions include the following:

- Hyphenated word are usually separated, and hyphenated numbers usually form one token.

- Sequences of numbers separated by slashes (e.g. in dates) form one token.

- Sequences that look like urls or email addresses form one token.

2.3 Filtering

We attempted to filter out all tokens that do not belong in English word n-gram counts. This includes tokens with any of the following characteristics:

- Malformed UTF-8 encoding.

- Tokens that are too long.

- Tokens containing any non-Latin scripts (e.g. Chinese ideographs).

- Tokens containing ASCII control characters.

- Tokens containing non-ASCII digit, punctuation, or space characters.

- Tokens containing too many non-ASCII letters (e.g. accented letters).

- Tokens made up of a combination of letters, punctuation, and/or digits that does not seem useful.

2.4 The Token "<UNK>"

All filtered tokens, as well as tokens that fell beneath the word frequency cutoff (see 3.1 below), were mapped to the special token "<UNK>" (for "unknown word").

2.5 Sentence Boundaries

Sentence boundaries were automatically detected. The beginning of a sentence was marked with "<S>", the end of a sentence was marked with "</S>". The inserted tokens "<S>" and "</S>" were counted like other words and appear in the n-gram tables. So, for example, the unigram count for "<S>" is equal to the number of sentences into which the training corpus was divided.

3. Frequency Cutoffs

3.1 Word Frequency Cutoff

All tokens (words, numbers, and punctuation) appearing 200 times or more (1 in 5 billion) were kept and appear in the n-gram tables. Tokens with lower counts were mapped to the special token "<UNK>".

3.2 N-gram Frequency Cutoff

N-grams appearing 40 times or more (1 in 25 billion) were kept, and appear in the n-gram tables. All n-grams with lower counts were discarded.

4. Data Format

4.1 Contents of Top-level Directory

Directory "doc": documentation (replicated on dvd1 - dvd5).

Directory "data": n-gram data.

4.2 Contents of "data" Directory

One sub-directory per n-gram order: "1gms", "2gms", "3gms", "4gms", "5gms".

4.3 Contents of "1gms" Subdirectory (on dvd1)

The "1gms" subdirectory contains the unigram information.

4.3.1 File "vocab.gz"

The file "vocab.gz" contains the vocabulary sorted by word in unix sort-order. Each word is on its own line:

```
WORD <tab> COUNT
```

The file is compressed using gzip.

4.3.2 File "vocab_cs.gz"

The file "vocab_cs.gz" contains the same data as "vocab.gz", but sorted by count. Also gzip'ed.

4.3.3 File "total"

The file "total" contains the total token count of the corpus.

4.4 Contents of sub-directories "2gms", "3gms", "4gms", "5gms"

The subdirectories with the remaining n-gram counts are distributed over dvd1 - dvd5.

4.4.1 Files Ngm-KKKK.gz

The n-gram counts are stored in files named "Ngm-KKKK.gz", where N is the order of the n-grams, and KKKK is the zero-padded number of the file.

Each file contains 10 million n-gram entries. N-grams are unix-sorted. The files are gzip'ed. Each n-gram occupies one line:

```
WORD_1 <space> WORD_2 <space> ... WORD_N <tab> COUNT
```

4.4.2 File "Ngm.idx"

An index to which n-gram counts are contained in which n-gram count file is stored in the files "Ngm.idx", where N is the order of the n-grams. The index file contains one line for each n-gram file:

```
FILENAME <tab> FIRST_NGRAM_IN_FILE
```

5. Data Sizes

File sizes: approx. 24 GB compressed (gzip'ed) text files

Number of tokens:	1,024,908,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

6. Acknowledging the Data

We are very pleased to be able to release this data, and we hope that many groups find it useful in their work. If you use this data, we would like to ask you to acknowledge it in your presentations and publications. We are also interested in hearing what uses this data finds, so we would appreciate hearing from you how you used the data.

7. Contact Information

We would welcome comments, suggestions, questions about the contents of this data, suggestions for possible future data sets, and any other feedback. Please send email to: ngrams@google.com

Thorsten Brants, Alex Franz
Google Research
19-Apr-2006