機械学習の続きと PredicitonAPIの詳細

グループI4
2008MI233 鈴木健太
2008MI214 沢田天馬

目次

データマイニングと機械学習

データマイニングの例

ある商品が一般の消費者にどのように評価されているかを
Web上のデータから解析すること

Webサイト、ブログの
様々なデータ

解析

結果

自然言語処理：入力された文章をまず分析し、その後に文書全体の意味を捉える

- 文の抽出
- 形態素解析
- 構文解析
- 意味解析
- 談話理解

This is a pen. → This | is | a | pen

これは鉛筆です。 → これは | 鉛筆 | です

データの対象をテキストに絞ったものを
テキストマイニングという

n-gramによる特徴抽出

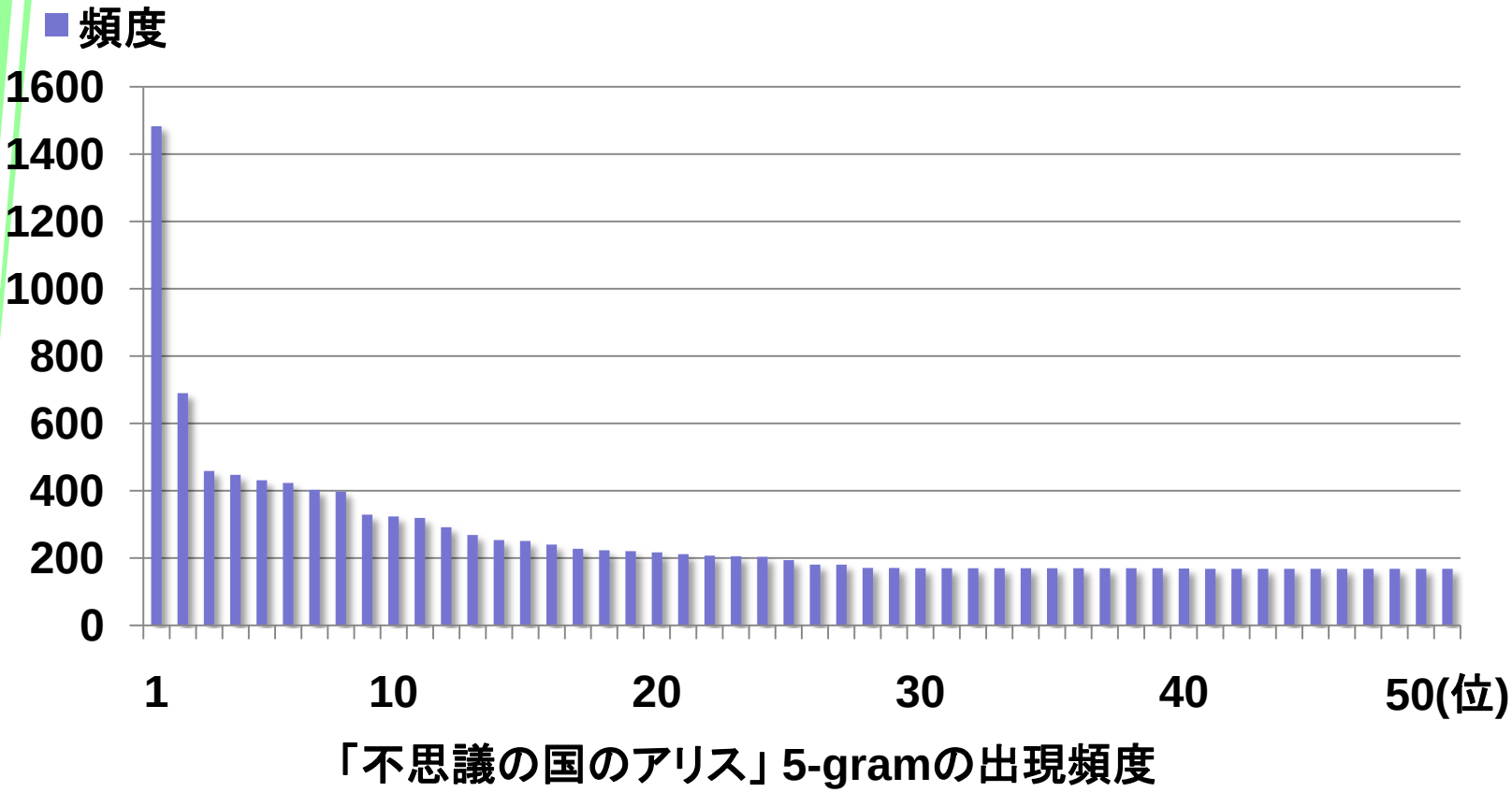
解析対象：This_is_a_pen.
3-gram　：Thi
his
is_
s_i
_is
…

図：n-gramの抽出

表：ルイス・キャロル作「不思議の国のアリス」
全文についての5-gramの出現頻度

出現順位	頻度	5-gram
1	1483	the
2	690	and
3	459	d the
4	447	said
5	431	, and
6	423	she
7	403	said
8	398	Alice

n-gramによる特徴抽出（2）



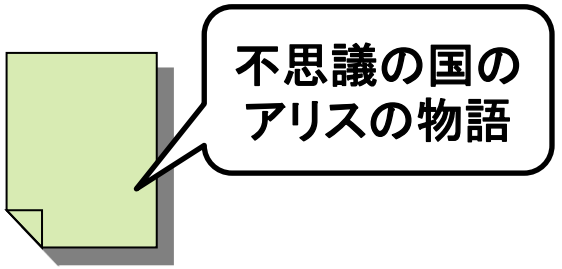
「不思議の国のアリス」5-gramの出現頻度

・頻度の高い5-gramは少数しかなく、
　　頻度の低い5-gramは数が多いという特徴がある

n-gramの出現頻度に基づく学習の実装

（1）テキストデータの入手

プロジェクトグーテンベルグによって公開
http://www.gutenberg.org/wiki/Main_Page



alice.txt

（2）n-gramの作成(ngram.c)

1文字ずつテキストを読み込みながら、過去n文字分のデータを出力

$ ./ngram 5 < alice.txt > alice5gram.txt

ここに結果を格納

$ cat alice5gram.txt

結果を表示

5字で分けられた
n-gram

n-gramの出現頻度に基づく学習の実装（2）

（3）n-gramの出現頻度の解析（rank.c）

n-gramから出現頻度表を作成するために、頻度を数え上げ、頻度順に並べるプログラムrank.cを作成する

収まりきらずに上位の結果が見れないのでlessコマンドを使いました

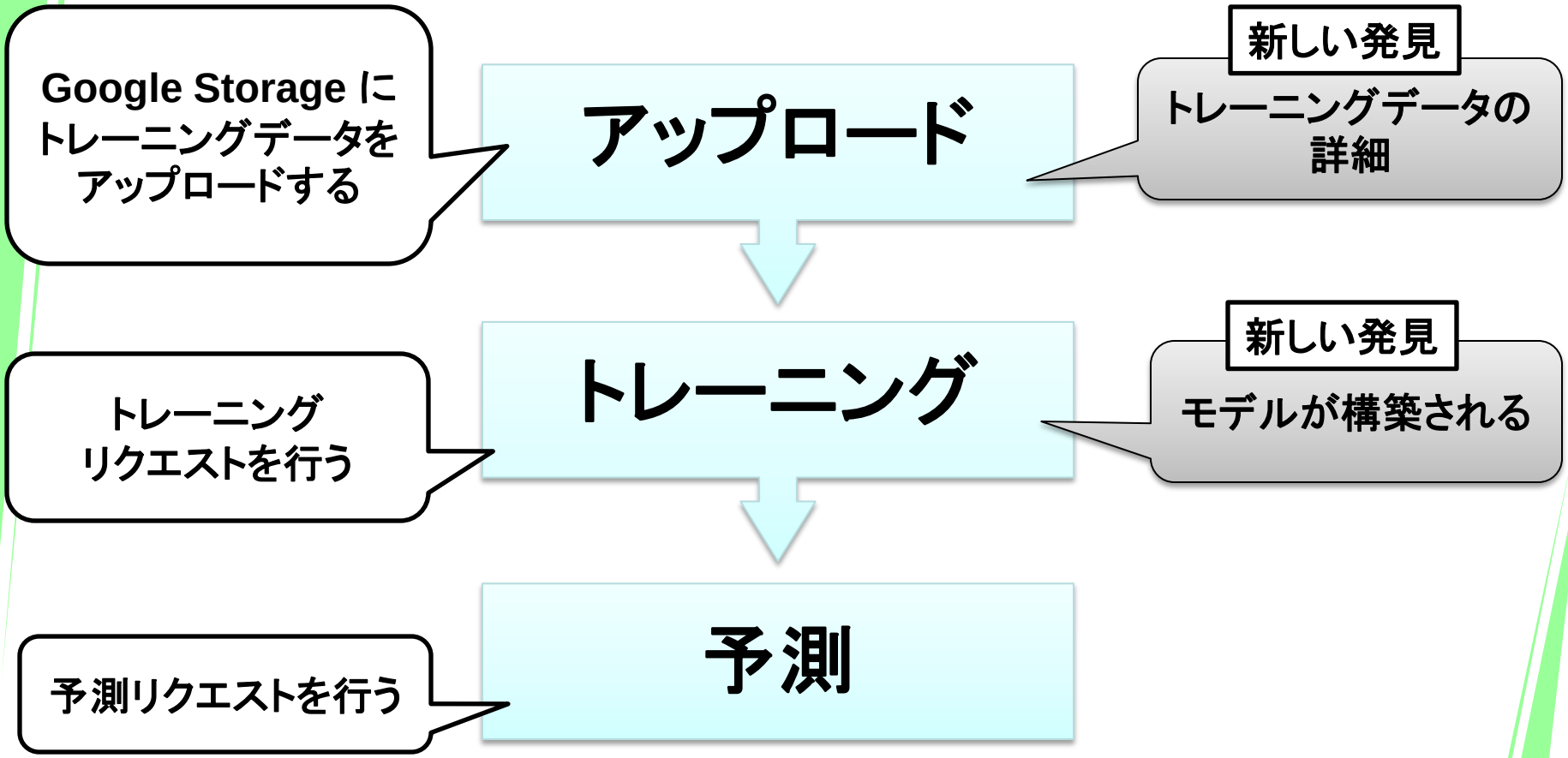
```
$ ./ngram 5 < alice.txt | ./rank | less
```

結果を表示

```
1483     the
690      and
459     d the
447      said
431     , and
423      she
403     said
398     Alice
329      Alic
324     ' sai
319
292      was
269      you
```

5字で分けられた
不思議の国のアリスの
n-gramの頻度順

PredictionAPIの復習

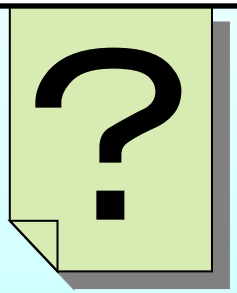


3つのステップでPredictionAPIの利用が可能

トレーニングデータの中身

前々回からの課題

language_id.txtの詳細



フランス語

M. de Troisvilles, comme s'appelait encore sa famille en Gascogne, ou M.

フランスの物語　三銃士の一章

スペイン語

En efecto, rematado ya su juicio, vino a dar en el

スペインの物語 ドンキホーテの一章

英語

And she went on planning to herself how she would manage it.

イギリスの物語 不思議の国のアリスの一章

この3つの物語の一章分がlanguage_id.txtに書かれていた

language_id.txtの中身

language_id.txt - メモ帳

ファイル(F) 編集(E) 書式(O) 表示(V) ヘルプ(H)

"French", "M. de Troisvilles, comme s'appelait encore sa fami
de Tr駈ille, comme il avait fini par s'appeler lui-m麥e ・Par
commenc・comme d'Artagnan, c'est-・dire sans un sou vaillant,
d'audace, d'esprit et d'entendement qui fait que le plus pauv
re輟it souvent plus en ses esp駸ances de l'h駸itage paternel
gentilhomme p駸igourdin ou berrichon ne re輟it en r饌lit・ Sa
son bonheur plus insolent encore dans un temps o・les coups p
l'avaient hiss・au sommet de cette 頡helle difficile qu'on ap
cour, et dont il avait escalad・quatre ・quatre les 頡helons.
"English", "Et, saluant la dame d'un signe de t黍e, il s'駉an
tandis que le cocher du carrosse fouettait vigoureusement son
interlocuteurs partirent donc au galop, s'駉oignant chacun pa

文字化けをおこしているが、3つの物語が書かれている

適当ではなく、トレーニングデータにも形式があることがわかった

トレーニングデータの仕組み

Comma-separated value format(CSV)　カンマ区切りモデル

トレーニングデータ

“English” , “The quick brown fox jumped over the lazy dog”

“English” , “To err is human, but to really foul things up you need a computer”

“Spanish” , “No hay mal que por bien no venga”

“Spanish” , “La fe mueve vencida”

出力となる部分

自身で作成可能

トレーニングデータ内にない未知の文章

La fe mueve montanas

To be or not to be, that is the question

トレーニングデータと比べた判断

Prediciton APIの理解

機械学習アルゴリズムについての修正

Googleの提供する
複数ある機械学習アルゴリズムの中から最適なものを選んでいる

別の使い道の発見

二つのトレーニング後データのモデルが存在

分類モデル

- 分類正確度
- 分類分けが可能

回帰モデル

- 平均二乗誤差
- 回帰分析が可能

Hello Predictionで使用

新しく発見したモデル

今後の課題

日本語のテキストデータを用いた例題を実装

機械学習に関する知識をもう少しつける

PredictionAPIの理解が深まってきたので、APIの利用方法を考える

PredictionAPIを用いてサービスの連携

参考文献

- **Google I/O 2010 - BigQuery and Prediction APIs**
  http://www.youtube.com/watch?v=dbkwv1wjs3A

- **はじめての機械学習　小高　知宏**

- **プロジェクトグーテンベルグ　http://www.gutenberg.org/wiki/Main_Page**