found it project セミナー#1

Pythonと機械学習を使った文章の内容予測

辻 真吾（www.tsjshg.info）

2016.10.28

自己紹介

- 辻 真吾（つじ しんご）1975年生まれ
- 都内のとある大学で研究職やっています
  - 専門：生命科学分野の大規模データ解析
- Pythonとデータサイエンス、機械学習
  - 昔は：C、C++、Javaなど
- メディア工房のプロジェクトに協力しています

自然言語処理

- 人が使う言語を計算機に理解させるための技術
- 人工知能研究の黎明期（1950年代）からある
- そんな簡単じゃない

1. 職場の上司とW不倫中です。
2. 職場の上司と付き合っています。私は既婚者です。

AIブームの背景

- データの爆発的増加
  - 計算機の高性能化とネットワークの発達
- 機械学習アルゴリズムの発展
  - Deep Learningの登場と進化

自然言語処理の世界では

RやSASも悪く無いけど、データサイエンスならPythonがいいらしい。

Deep Learningでは、世界的に見て、Pythonの独壇場。

Rはたしかに入門しやすいけど、汎用言語のPythonなら、
　データ解析以外にも使える。

・
・
・

- 1つ1つの文章の意味をきちんと理解する

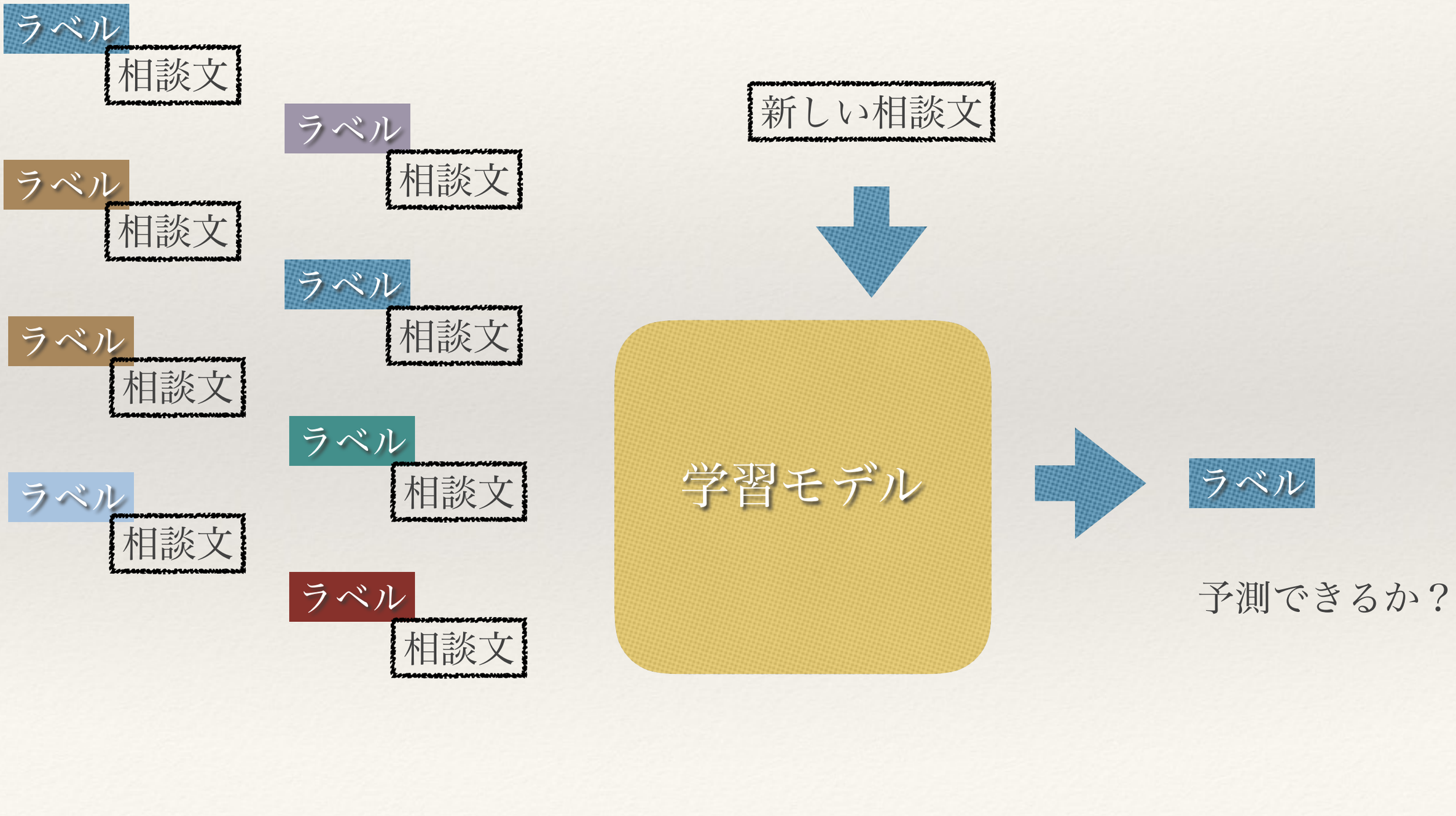
- 沢山の文章をまとめて処理して、なんとなく意味をとる

たとえば・・・

ユーザーさまからのお悩み相談
申し訳ありませんが、公開は控えさせてください。

これは「片想い」の相談（ユーザーの入力）

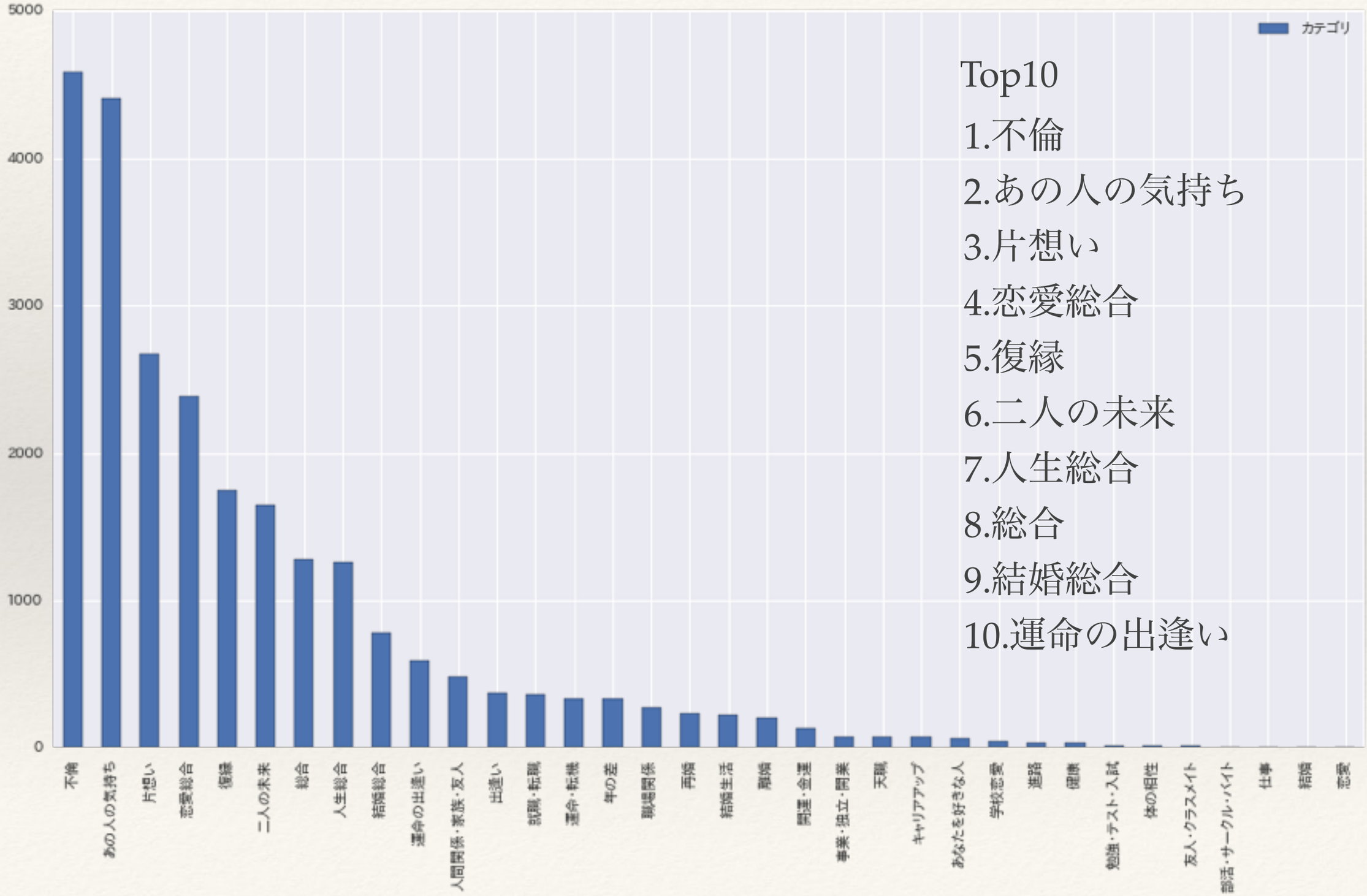
大量のデータを使った学習



利用したデータ

- メディア工房社内のデータから、約3万件のラベル付き相談文のデータで、学習モデルを作成
- 別の相談文を入力にして、ラベルを予測

データの全体像



Top10

1.不倫

2.あの人の気持ち

3.片想い

4.恋愛総合

5.復縁

6.二人の未来

7.人生総合

8.総合

9.結婚総合

10.運命の出逢い

解析の手順

1. MeCabを使った形態素解析

2. 単語辞書の作成（以降4まで、gensimを利用）

3. tf-idfを使った各相談文のベクトル表現

4. LSI（Latent Semantic Index）を使った次元縮約

5. Random Forestsを利用した予測モデルの構築

形態素解析

- MeCab、JUMAN++（京都大学）、JMAT（Justsystem）など
- すももももももものうち
- すもも / も / もも / も / もも / の / うち
- MeCab、JUMAN++ともPythonバインディングがあって便利

単語の頻度を計測

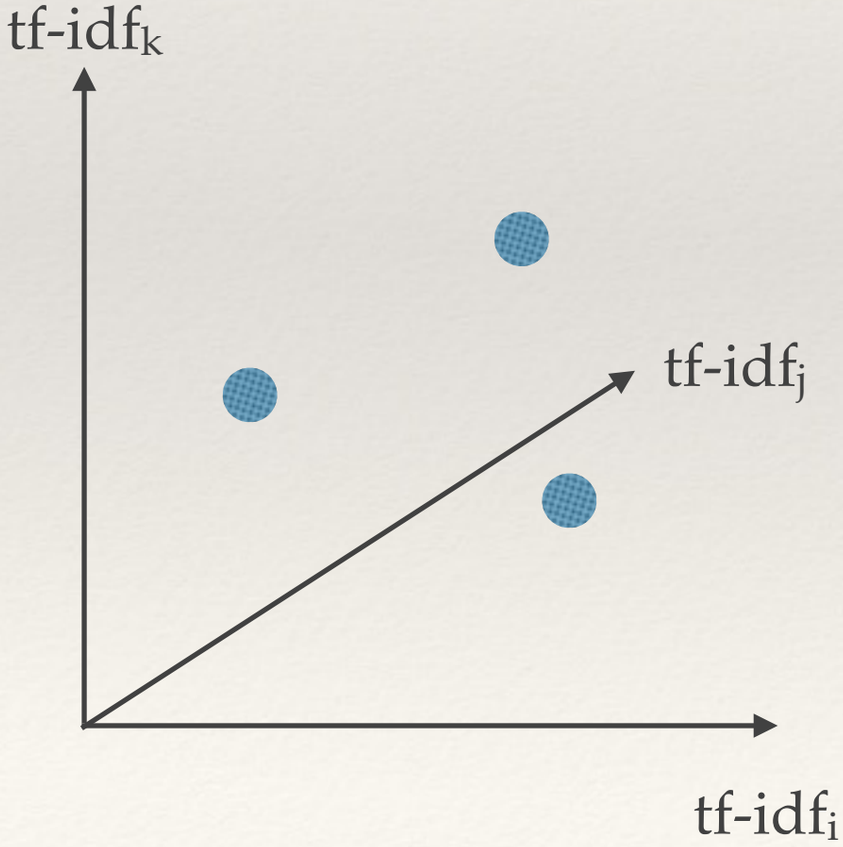
- Pythonは最高。
- 明日もPython、明後日もPython。
- 「は」、「も」は頻出なので省いて、辞書を作る
- 0:Python、1:最高、2:明日、3:明後日
- Pythonは最高。 → [(0, 1), (1, 1)]
- 明日もPython、明後日もPython。→ [(0, 2), (2, 1), (3, 1)]

tf-idf

- V:すべての単語数、N:すべての文書数
- t_{ij}:単語iが文書jに出現する頻度
  - ある文書に何回その単語が出てくるか？
- df_i:単語iが出てくる文書の数
  - 文書全体で、よく使われている単語かそうでないか？
- $\text{tf-idf}(i, j) = t_{ij} \log(N / df_i)$
  - 単語iが文書全体で稀な単語なのに、文書jに出てくるということは、その文書を特徴付ける良い指標と考える

潜在意味解析

- Latent Semantic Index
- 文書を、単語のベクトル空間で考える



トピックモデル

単語は多いし意味が重複することもあるので、この空間の次元数を縮約する。

22,000（単語） → 100〜200

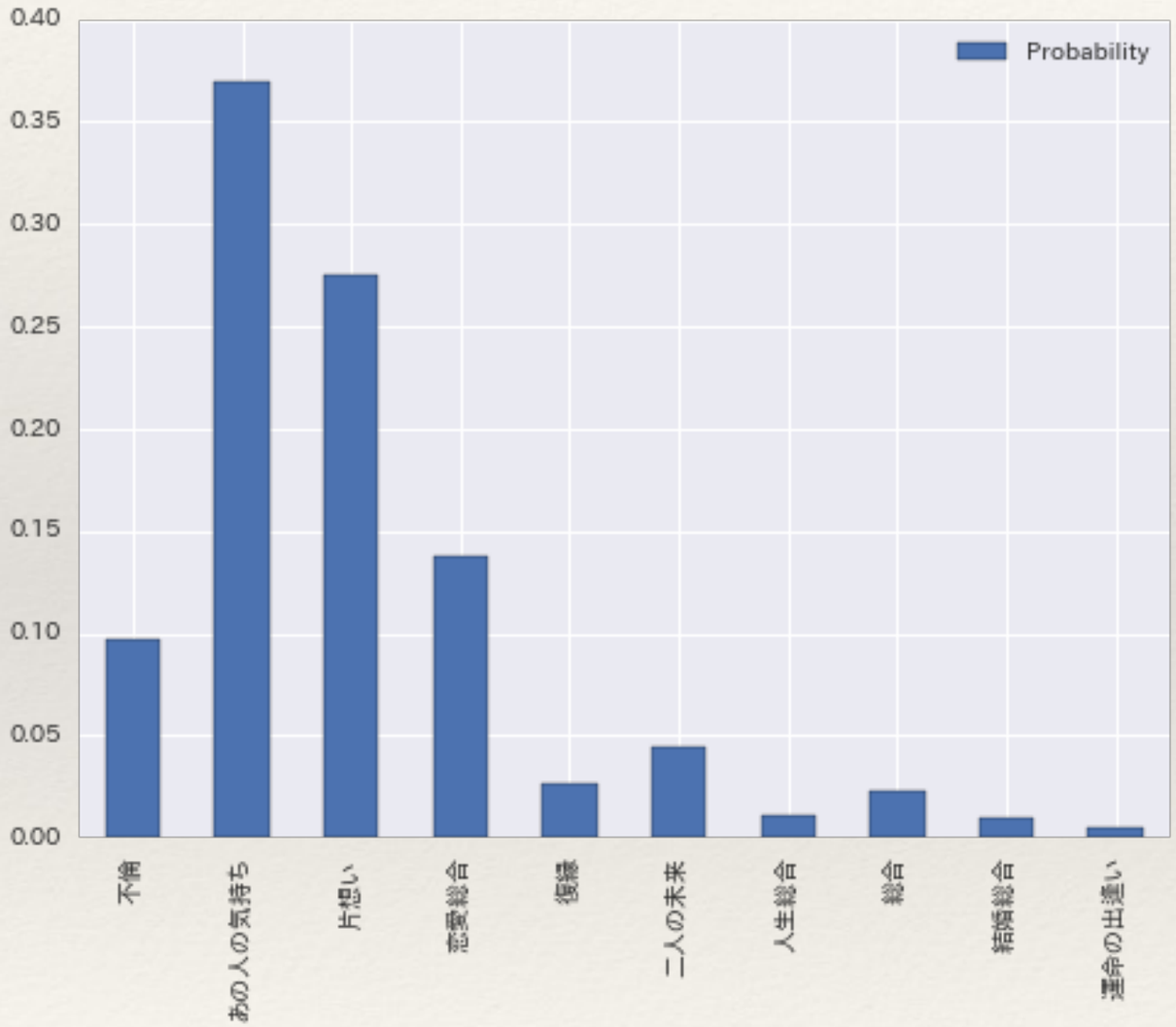
実際のデータ

	カテゴリ	相談内容	words	corpus	tfidf_corpus	lsi
0	片想い	[redacted]	[私, 歳, 年下, 男性, こと, 気, なる, いる, 彼, 同じ, 会社, 人出, 別...	[(4, 1), (9, 1), (18, 3), (19, 1), (32, 2), (6...	[(4, 0.01736358445943663), (9, 0.1779994609681...	[(0, 0.204106957421), (1, -0.00556939362687), ...
1	復縁		[私, 年, 前, 離婚, 子供, 人, いる, 不倫, いる, 彼, もう一度, 関係, ...	[(9, 3), (14, 1), (18, 4), (19, 2), (25, 4), (...	[(9, 0.23837370222342424), (14, 0.036571449732...	[(0, 0.209279709011), (1, 0.097054134537), (2,...
2	二人の未来		[私, 独身, 今, 好きな人, いる, その, 男性, 年, くらい, 前, イベント, ...	[(8, 1), (15, 1), (18, 5), (19, 1), (27, 1), (...	[(8, 0.044104145896312036), (15, 0.03885796471...	[(0, 0.214629324226), (1, -0.00517319103862), ...
3	出逢い		[私, 次, 誕生日, なる, 会社, 人, とか, 周り, 恋人, いる, 人, ばかり,...	[(1, 2), (4, 1), (18, 4), (32, 1), (34, 1), (5...	[(1, 0.07455083026877266), (4, 0.0112013137612...	[(0, 0.2069981888), (1, -0.141496672918), (2, ...
4	運命の出逢い		[半年, 付き合う, いる, 元彼, ずるずる, 復縁, わけ, また, 半年, 以上, 過...	[(4, 1), (8, 1), (18, 3), (52, 1), (54, 2), (5...	[(4, 0.010310417311352843), (8, 0.037688461301...	[(0, 0.181489594612), (1, -0.0782239526405), (...

解析結果

片想い

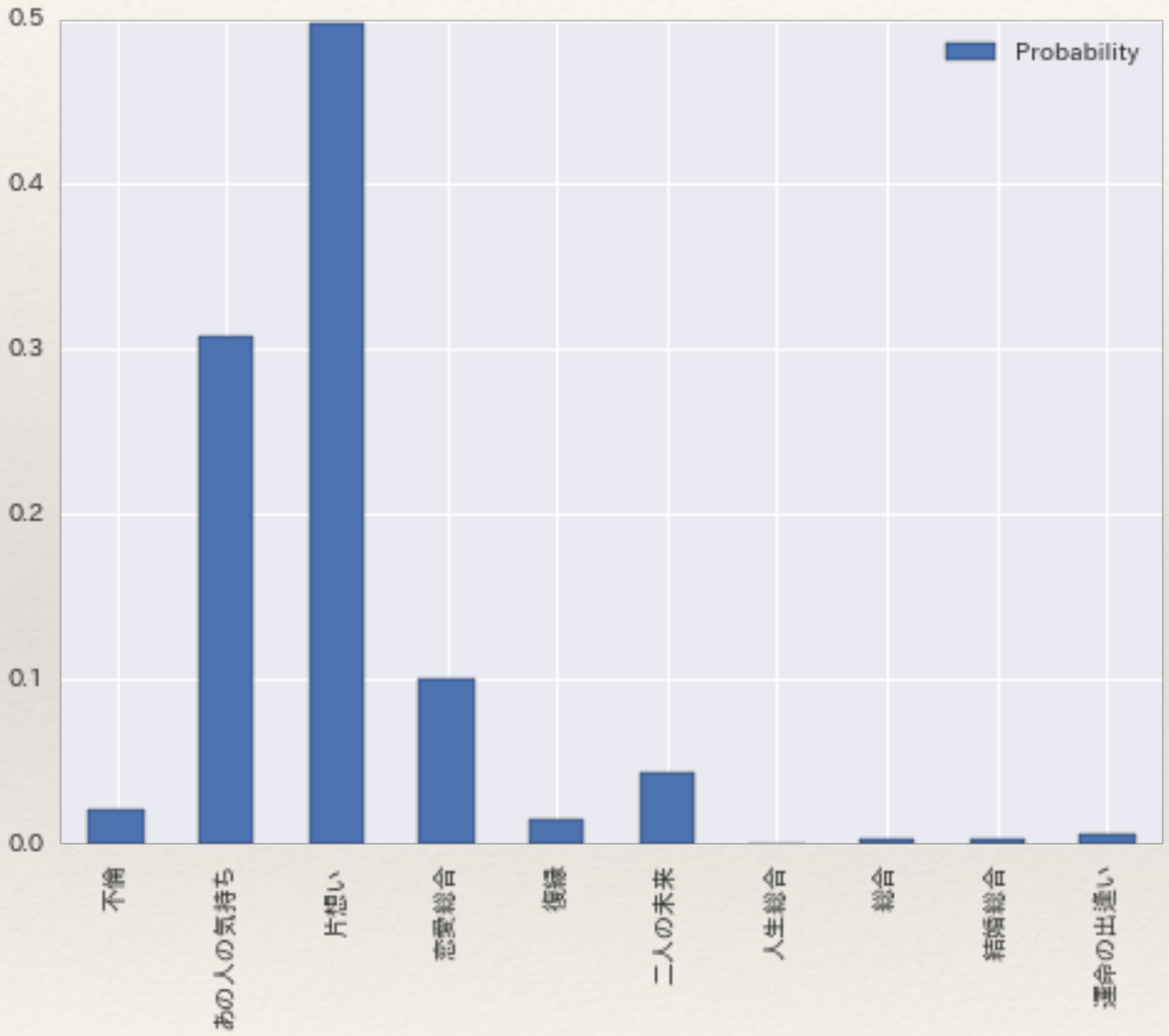
予測



あの人の気持ちが気になる片想い。
ちょっと不倫の要素も。

?

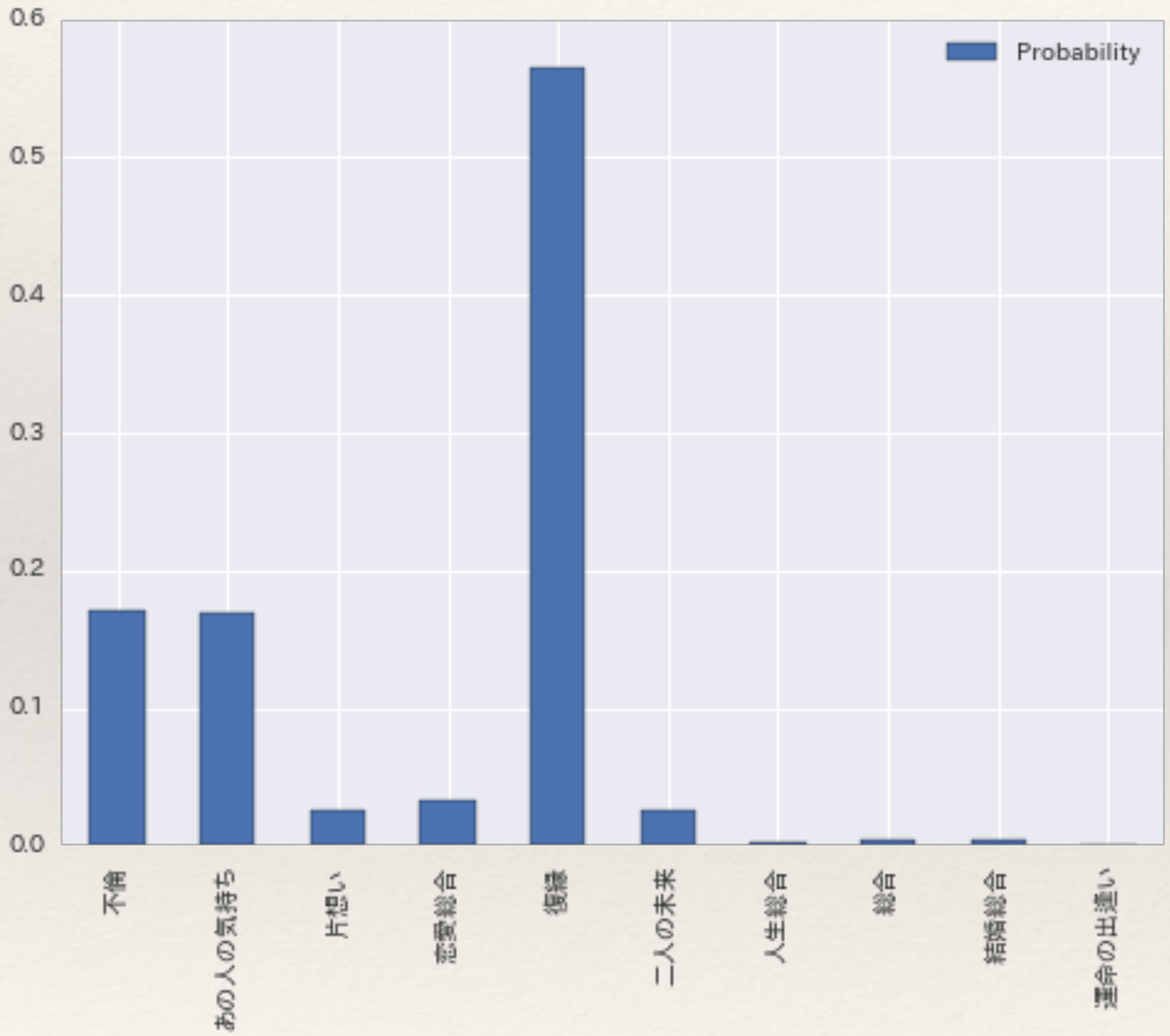
片想い



あの人の気持ちが気になる片想い。
不倫要素は低い。

?

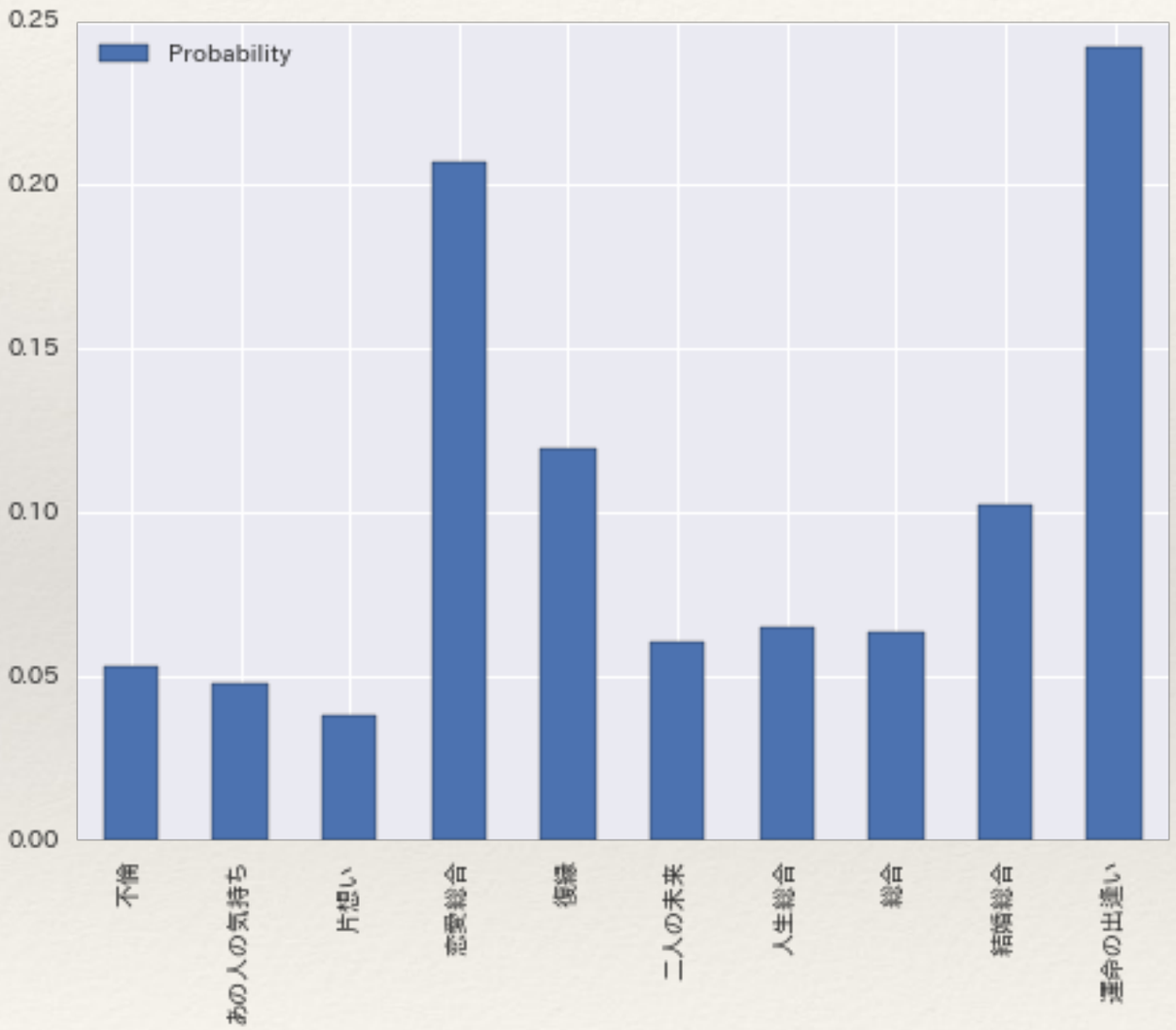
復縁



不倫要素も的確に予測

最後の問題です

運命の出逢い



いろいろとこれまでの経緯が書いてあって、迷うところではあるが、運命の出逢いで正解。

まとめ

- 最近の自然言語処理は力尽く
  - データ量の多さと計算機の性能に頼る
- MeCabで形態素解析、gensimを使って文章を単語の特徴ベクトルで表現し、Random Forestsで分類
- 長文を1つの文書データとすることで、ラベルを予測することができる