P16:ノンパラメトリックベイズモデルを用いた文分割

株式会社Preferred Infrastructure
徳永拓之

問題意識

- 文の分割は多くの場合、ルールベースで処理される
- ノイズとなり得る悩ましい単語がある
  - モーニング娘。 スプツニ子！
- ウェブデータだとルールでうまく分割できないこともよくある

問題意識

- 文の分割は多くの場合、ルールベースで処理される
- 分割箇所の悩ましい文も見られる

問題意識2

- 文には切れ目があるのではなく、始点と終点があるのでは？←仮説
- 文の生成モデルが考えられるのでは？

この味が いいねと君が 言ったから 七月六日は サラダ記念日

文1

文2

結局、何がしたいのか？

- 長い文章を切ってから解析するのではなく、前の方から文章を読みつつどこからどこまでが文なのかを解析したい
- 今はNonparametric Bayesではなくてもいい気がしています

今後の計画1(教師あり)

- 文始点と文終点の2つの分類器を作る
- どの始点とどの終点のペアがよいのかを判断する手法を考える
  - 正解データを与えればここも学習ができそう

((この味が いいね)と君が 言ったから 七月六日は サラダ記念日)

なぜ教師なしでやりたいのか？

- 文という単位を数式で定義したい
  - 悩ましい例の解析をシステムがこうだよと言ってくれると迷わなくていい
    - しかし、システムを信仰するのではなく、応用を見据えて自分でデータを作るべきなのではという自問自答もある