



# 研究室生活

李 哲豪

田浦研究室 M2

2019/11/27



# 自己紹介

名前: 李 哲豪 / リ テツゴウ / Zhehao Li

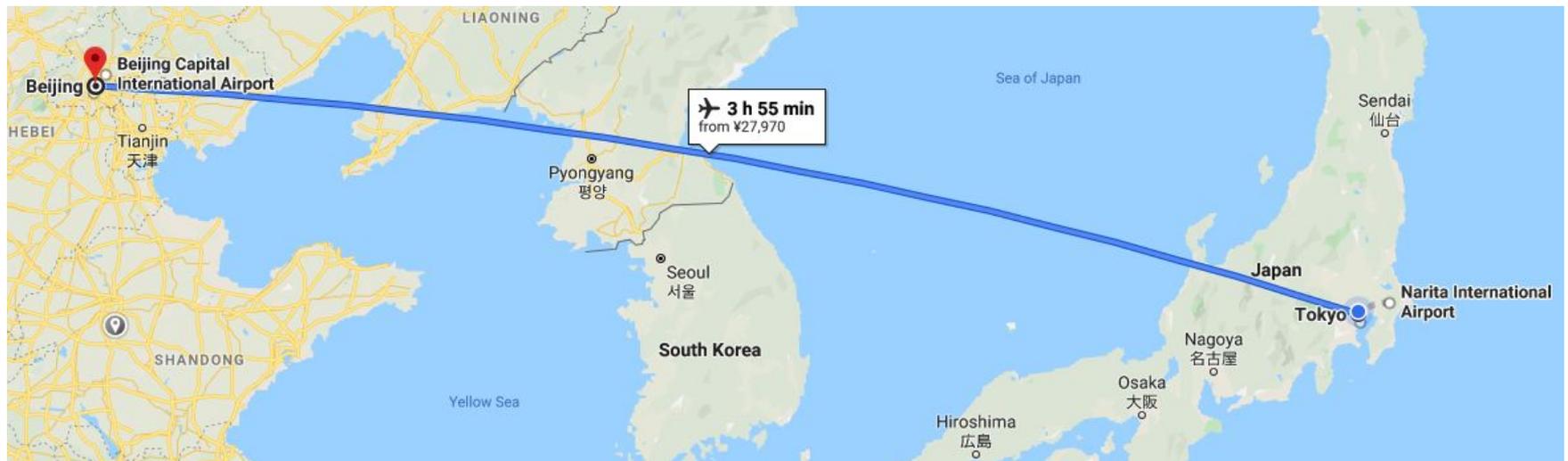
国籍: 中国

M2

学部: 北京大学 (Peking University)

2017年9月来日

2018年4月修士課程入学





# 日本に来る原因

- どこかに留学したい
- 日本語面白い！勉強したい！
- 近い
- その他...



# なぜ田浦研？

- 田浦先生は優しくて強いし、教育にも熱心
- 研究内容には興味がある
- 留学生も何人かがいる
- 英語と日本語両方も通じそう
- ...



# 研究室に入ったら...

- 研究テーマをやりたい分野に決められる
- 先生だけではなく、みなさんも優しい
- 英語を勉強できる環境
- ~~みんな強いのでストレスを感じる~~



# 自分の研究テーマ

- ドキュメントの類似検索、類似度計算
- 従来手法: vector space model
- ドキュメント→高次元vector



## Vector space model

- “This is a cat”  $\rightarrow [1, 1, 1, 1, 0]$
- “This is a dog”  $\rightarrow [1, 1, 1, 0, 1]$

this is a cat dog

- 類似度: cosine similarity, この場合は 0.75
- 問題点: 共通(同じ)単語だけを考慮する

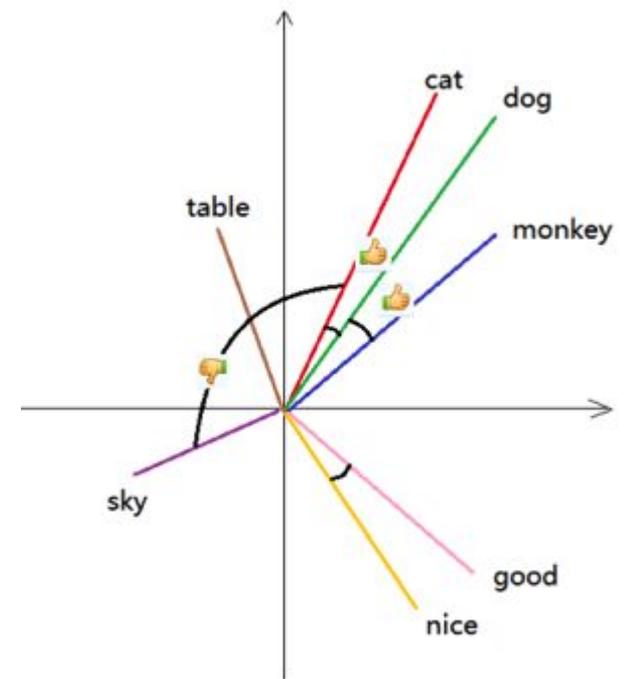
## 問題点

- 類似単語も考慮したい
- 例えば「computer」と「laptop」意味は近い
- 同じ単語ではないので類似度に貢献しない



# Word embedding

- 単語埋め込み (word embedding)
- 機械学習を用いた手法
- 単語 → 高次元vector
- 意味が近い単語のvectorは近い



Projection of the embedding vectors to 2-D



- Word embeddingを使ってドキュメントのvectorを生成
- その上にドキュメントの類似度を計算する
- 似ている単語も類似度に影響する





- どの手法でドキュメントのvectorを生成する
- ドキュメント類似検索の加速化
- ...



# みんなの研究テーマ

- 並列プログラミング、system softwareがメイン
- 自然言語処理(自分)、深層学習とかもある
- やりたいことがあれば先生と相談できる



# まとめ

- 先生と学生のみなさんは強くて優しい
- 色々勉強できる(研究、英語、日本語とか)
- 充実で楽しい生活ができる！
- ~~留学生の日本語(英語も可)先生になれる~~