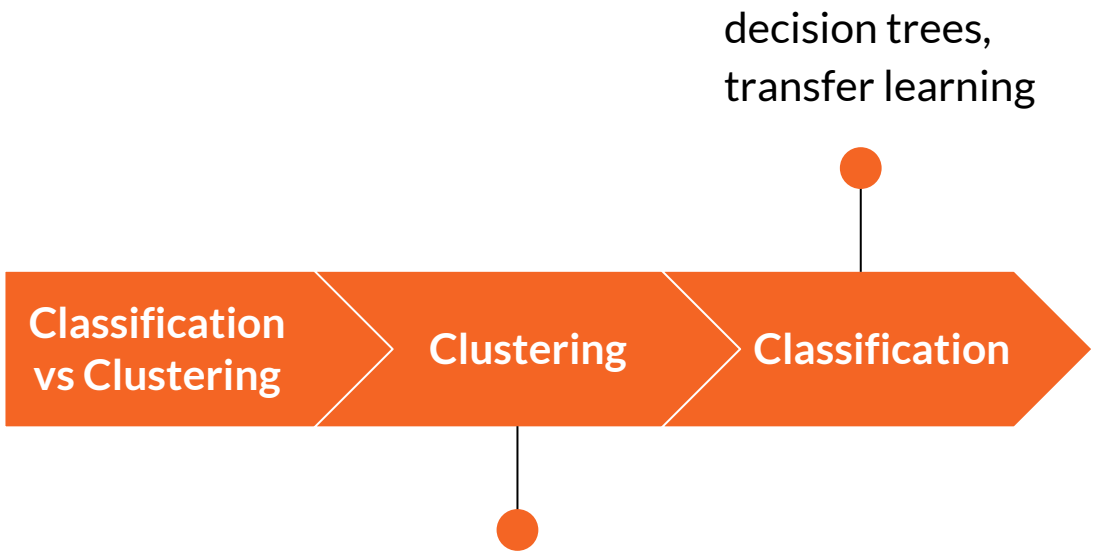
Clustering & Classification

Chris Swart Data Scientist @ Resolver

15/10/2018 - mcubed.london

Overview


decision trees,
transfer learning
Classification vs Clustering
Clustering
Classification
Topic modelling with LDA,
keyword extraction

Clustering vs Classification

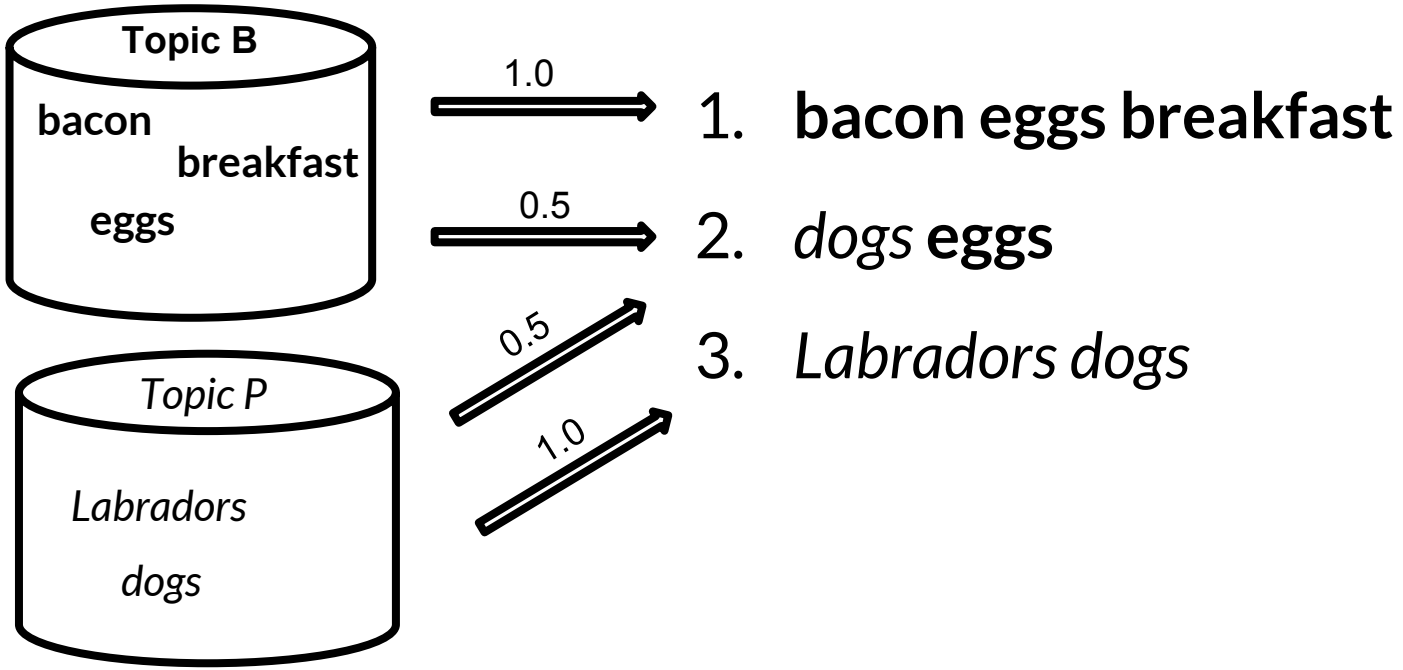
- Discover patterns based on **unlabelled** data
- **Easier to gather data** as we don't need labels
- Less clear how to interpret
- More flexible structure

- Utilise patterns based on a **labels** you have
- **More expensive** to gather data
- **Interpretable**

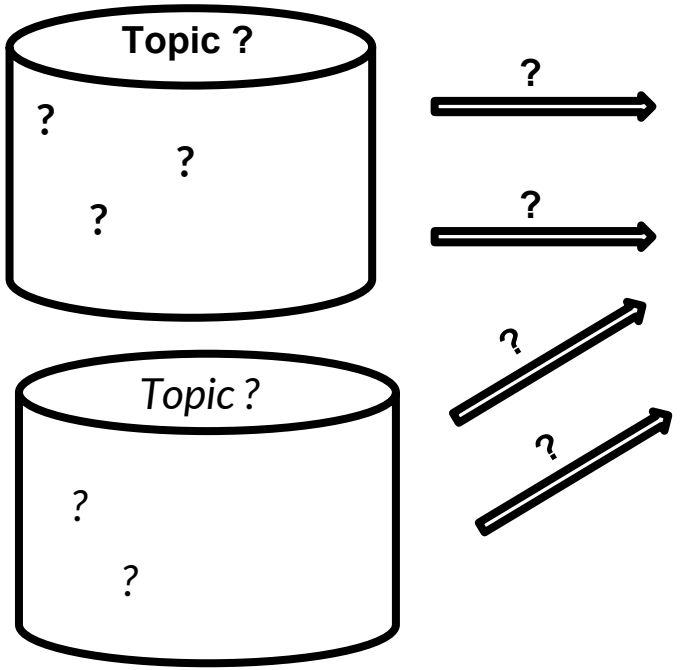
Topic modelling with LDA

1. I had **bacon** and **eggs** for **breakfast**.
2. *Labradors* are cute *dogs*.
3. My *dogs* do not like **eggs**.

Generative process with LDA

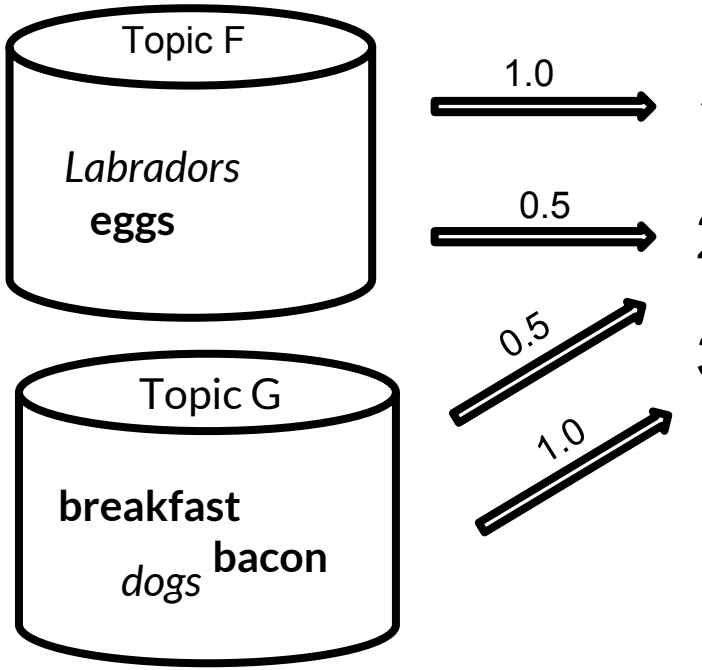


Inference with LDA



1. **bacon eggs breakfast**
2. *dogs* **eggs**
3. *Labradors dogs*

Generative process with LDA

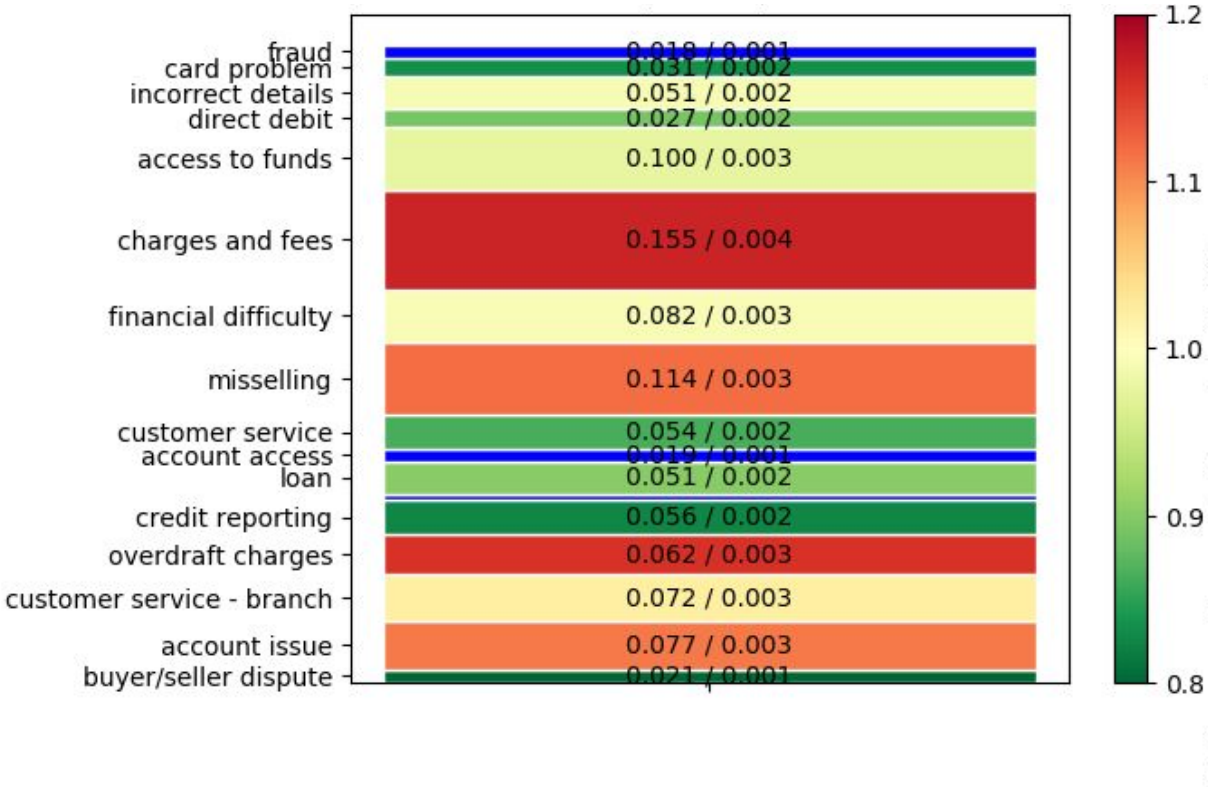


1. *Labradors* **eggs**
2. *Labradors* **breakfast**
3. *dogs* **bacon**

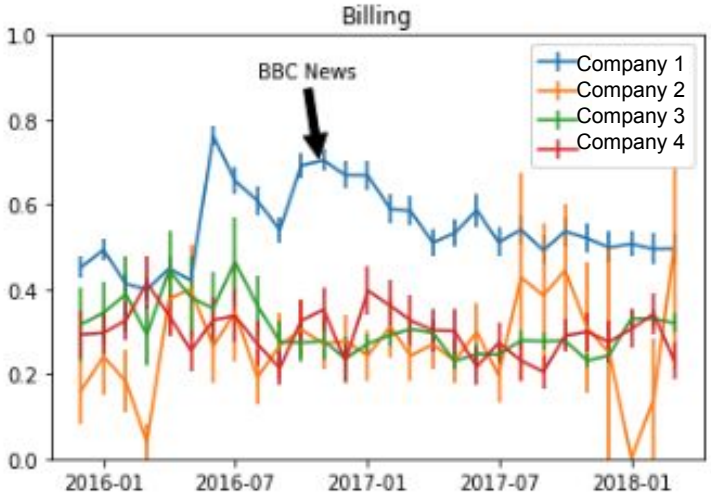
Evaluating LDA

- Word intrusion [eggs, breakfast, Labradors, bacon]
- Consolidate topics and annotate documents

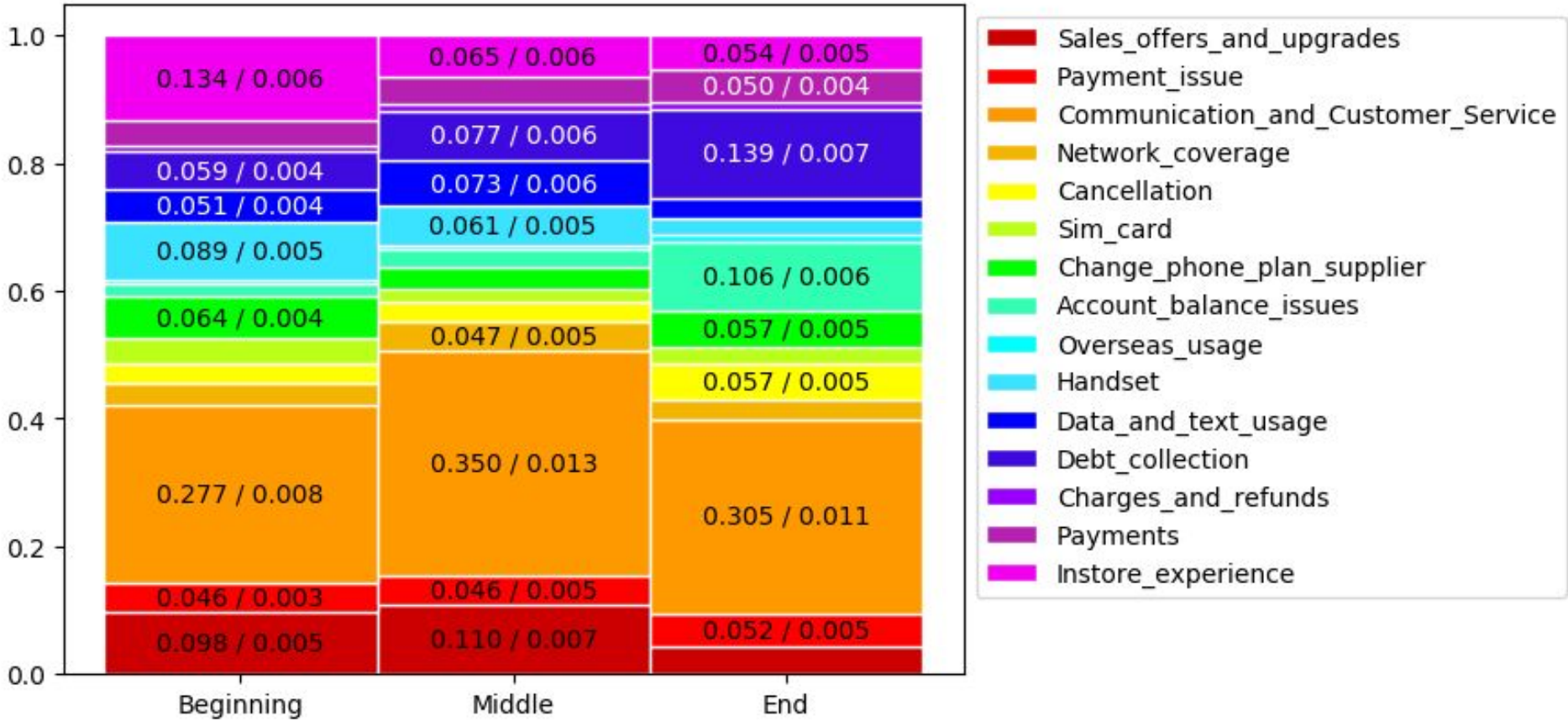
Bank1 performance compared to industry



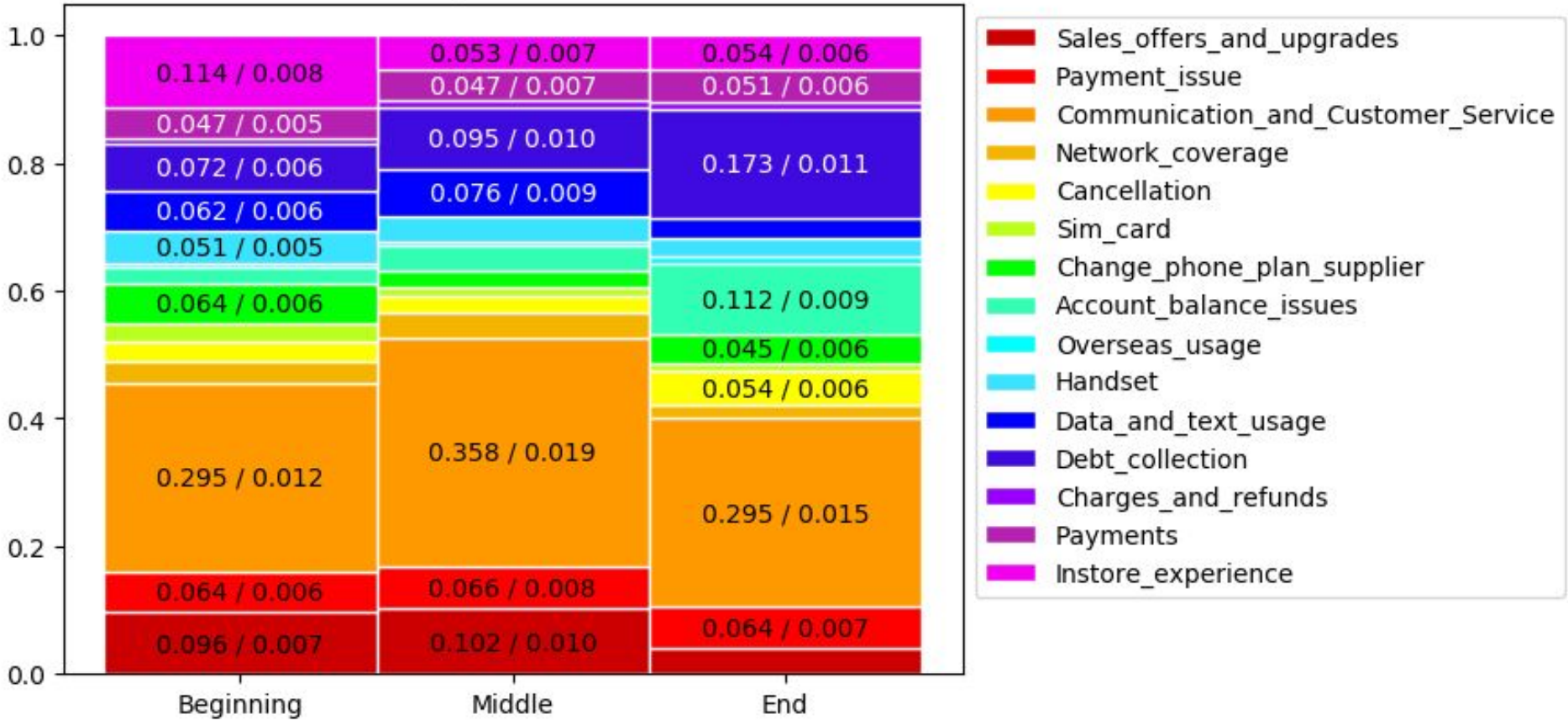
Case study - Telco1



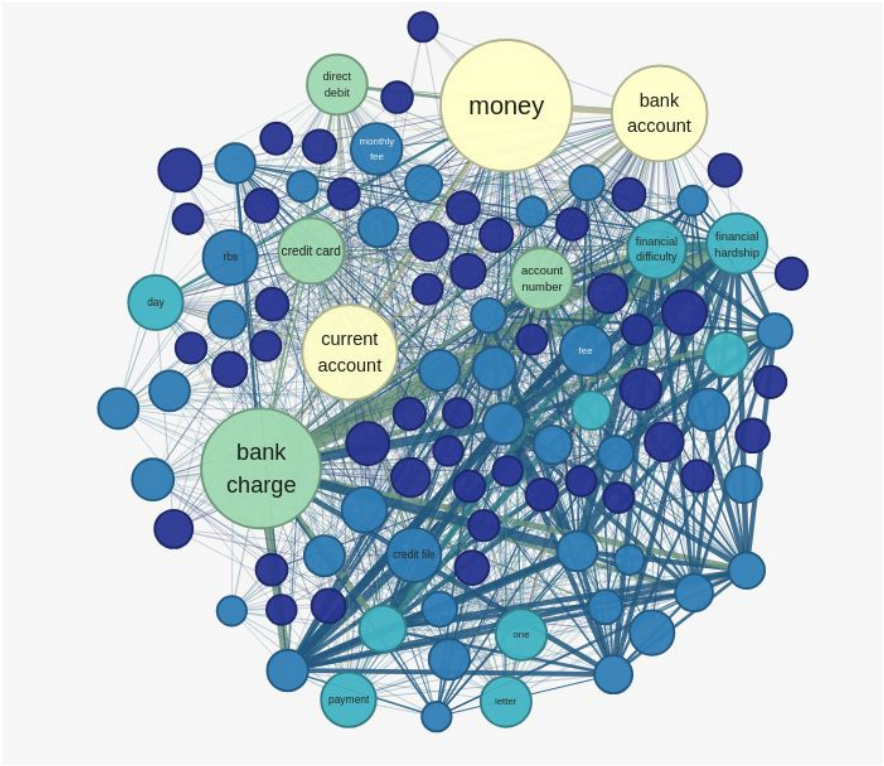
Industry user journey



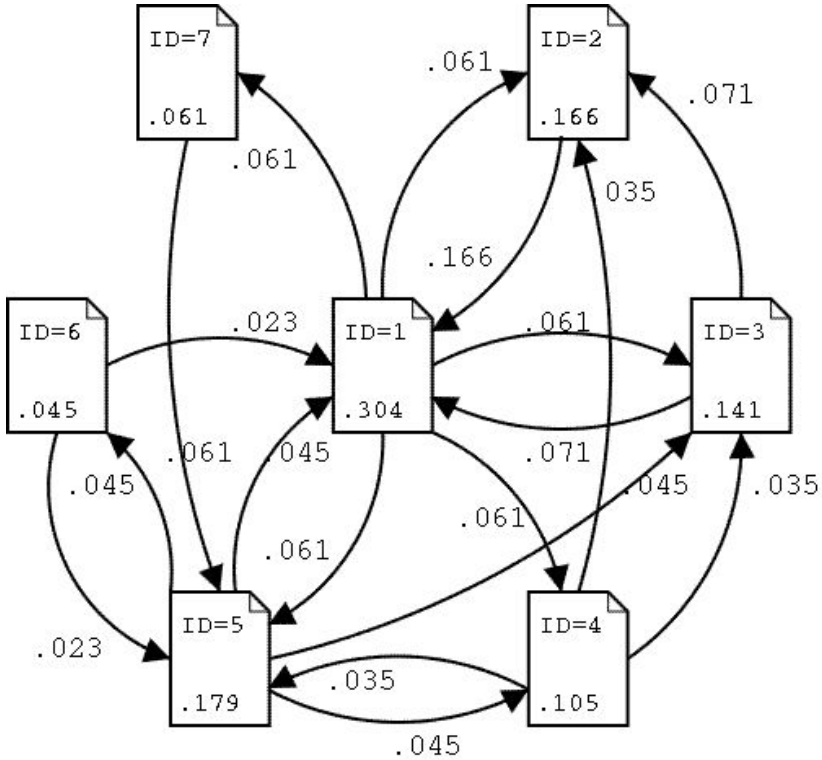
Telco1 user journey



Keyword extraction with TextRank



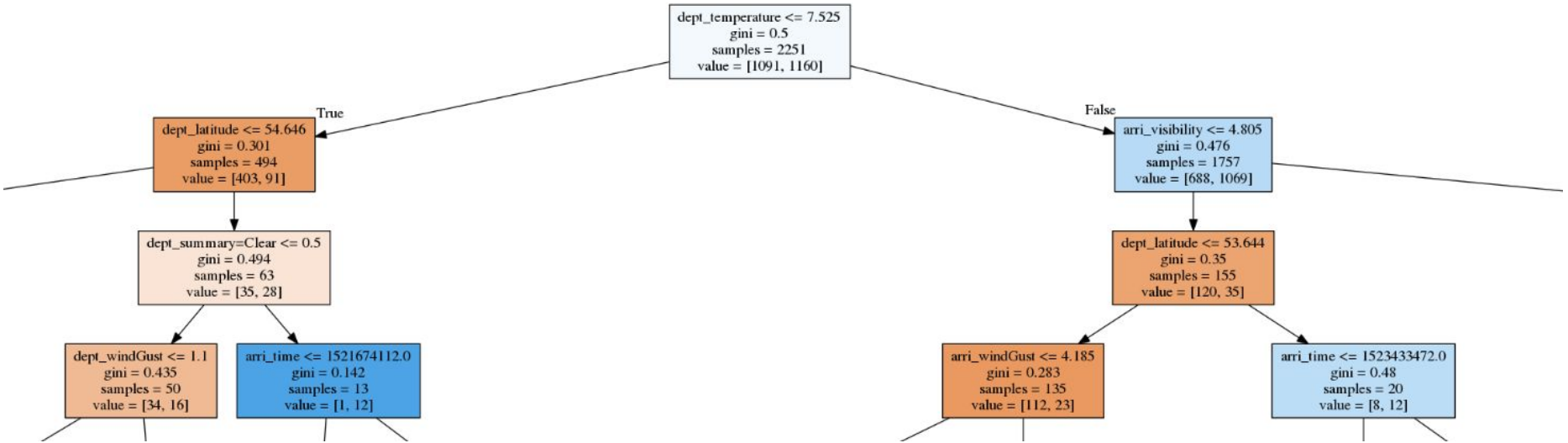
Pagerank + Text data = TextRank

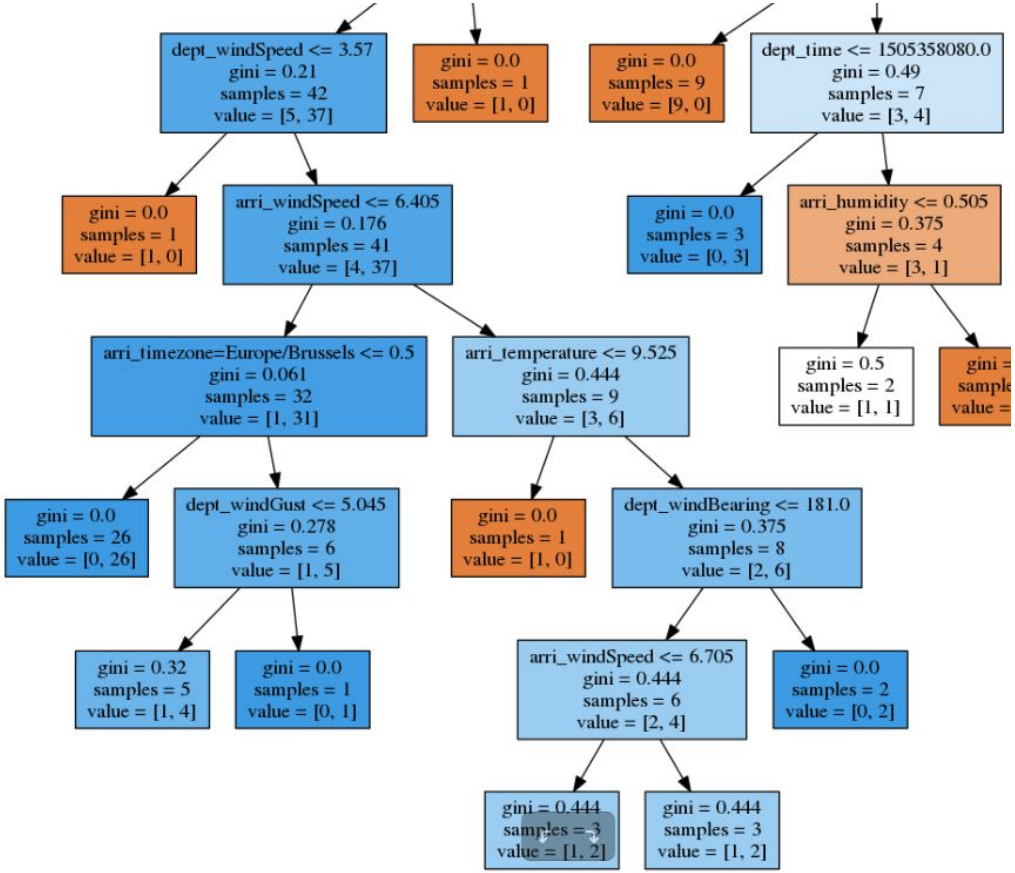


Evaluating keyword extraction

- Supervised ranking problem
- User feedback

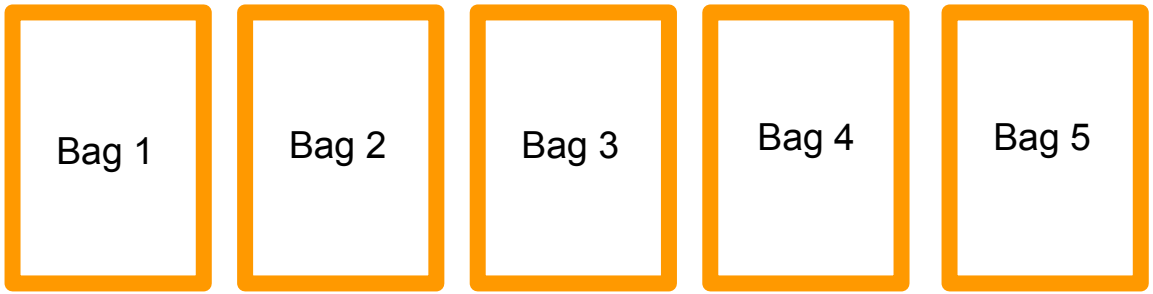
Decision trees




dept_windSpeed <= 3.57
gini = 0.21
samples = 42
value = [5, 37]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 9
value = [9, 0]
dept_time <= 1505358080.0
gini = 0.49
samples = 7
value = [3, 4]
gini = 0.0
samples = 1
value = [1, 0]
arri_windSpeed <= 6.405
gini = 0.176
samples = 41
value = [4, 37]
gini = 0.0
samples = 3
value = [0, 3]
arri_humidity <= 0.505
gini = 0.375
samples = 4
value = [3, 1]
arri_timezone=Europe/Brussels <= 0.5
gini = 0.061
samples = 32
value = [1, 31]
arri_temperature <= 9.525
gini = 0.444
samples = 9
value = [3, 6]
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 26
value = [0, 26]
dept_windGust <= 5.045
gini = 0.278
samples = 6
value = [1, 5]
gini = 0.0
samples = 1
value = [1, 0]
dept_windBearing <= 181.0
gini = 0.375
samples = 8
value = [2, 6]
gini = 0.32
samples = 5
value = [1, 4]
gini = 0.0
samples = 1
value = [0, 1]
arri_windSpeed <= 6.705
gini = 0.444
samples = 6
value = [2, 4]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.444
samples = 3
value = [1, 2]
gini = 0.444
samples = 3
value = [1, 2]

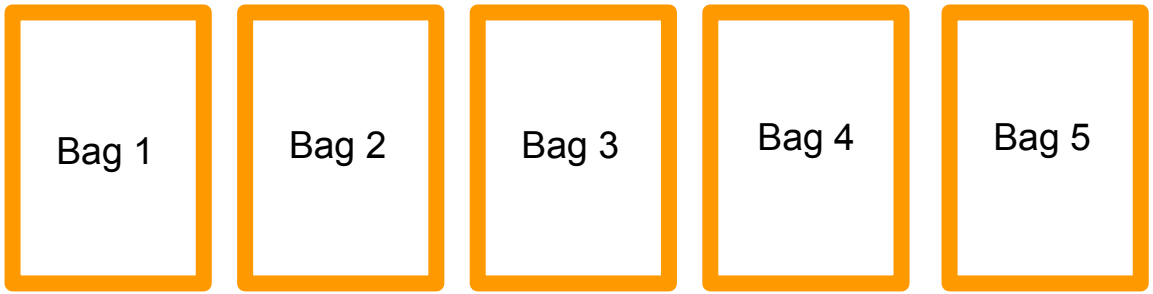
Decision forests

- Split the data
- Build a decision tree on each split
- Average their outputs



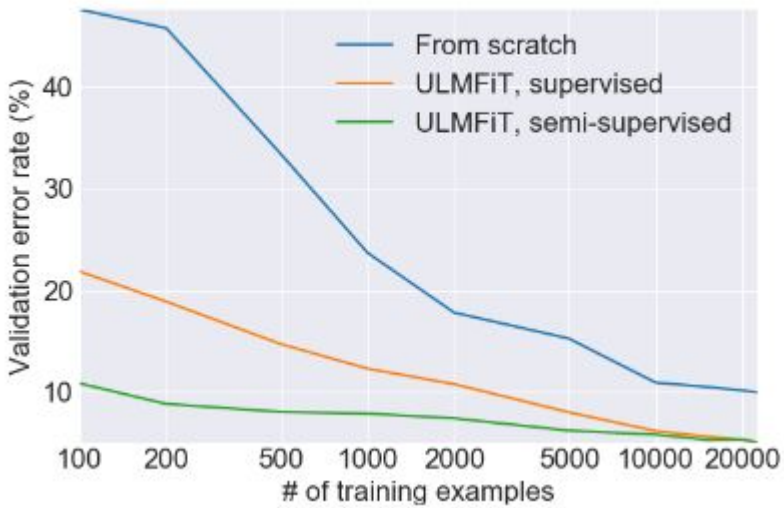
Evaluating decision forests

- Out of Bag scores
- Train / test / validation split

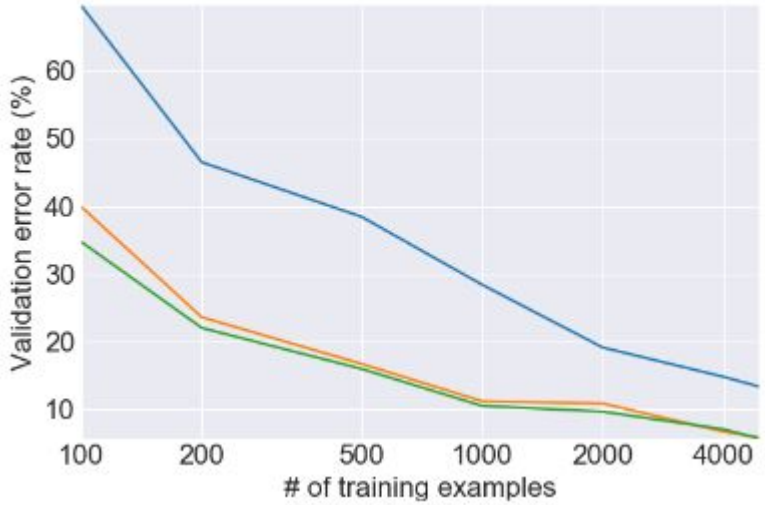


Transfer learning

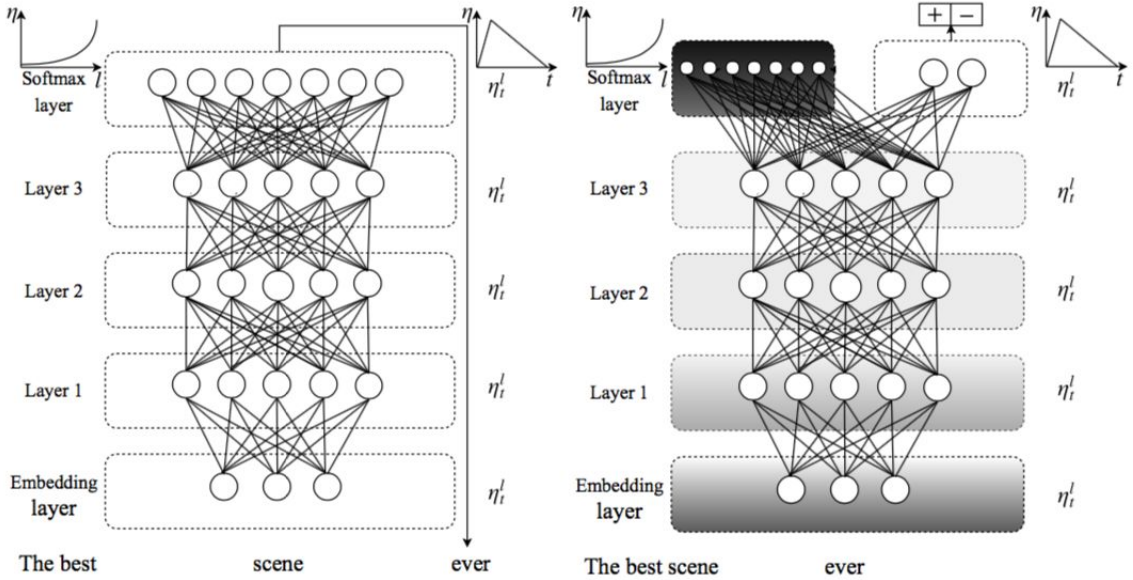
Movie reviews



Topic extraction of non-complaints

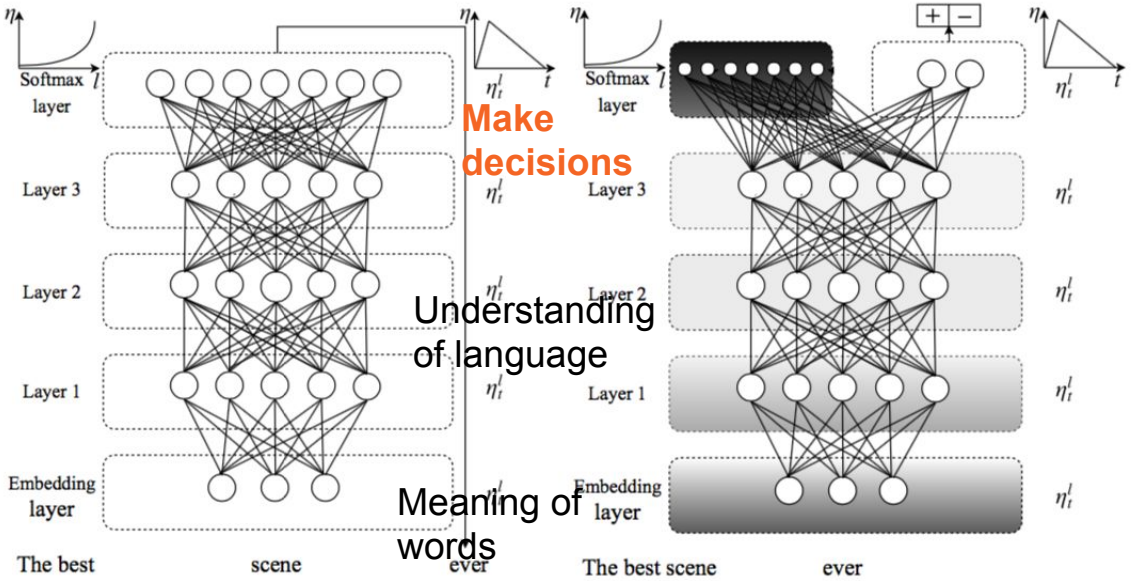


How does it work?



- The more knowledge the Neural Network has before it starts the less it has to work out.
- Here we train a “language model” first.

How does it work?



- The more knowledge the Neural Network has before it starts the less it has to work out.
- Here we train a “language model” first.

References

- Chang, Jonathan, et al. "Reading tea leaves: How humans interpret topic models." Advances in neural information processing systems. 2009.
- Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2018.

Questions

If you have any questions about the talk you can email me at chris@resolver.co.uk or you can message me on twitter swartchris8

Thanks