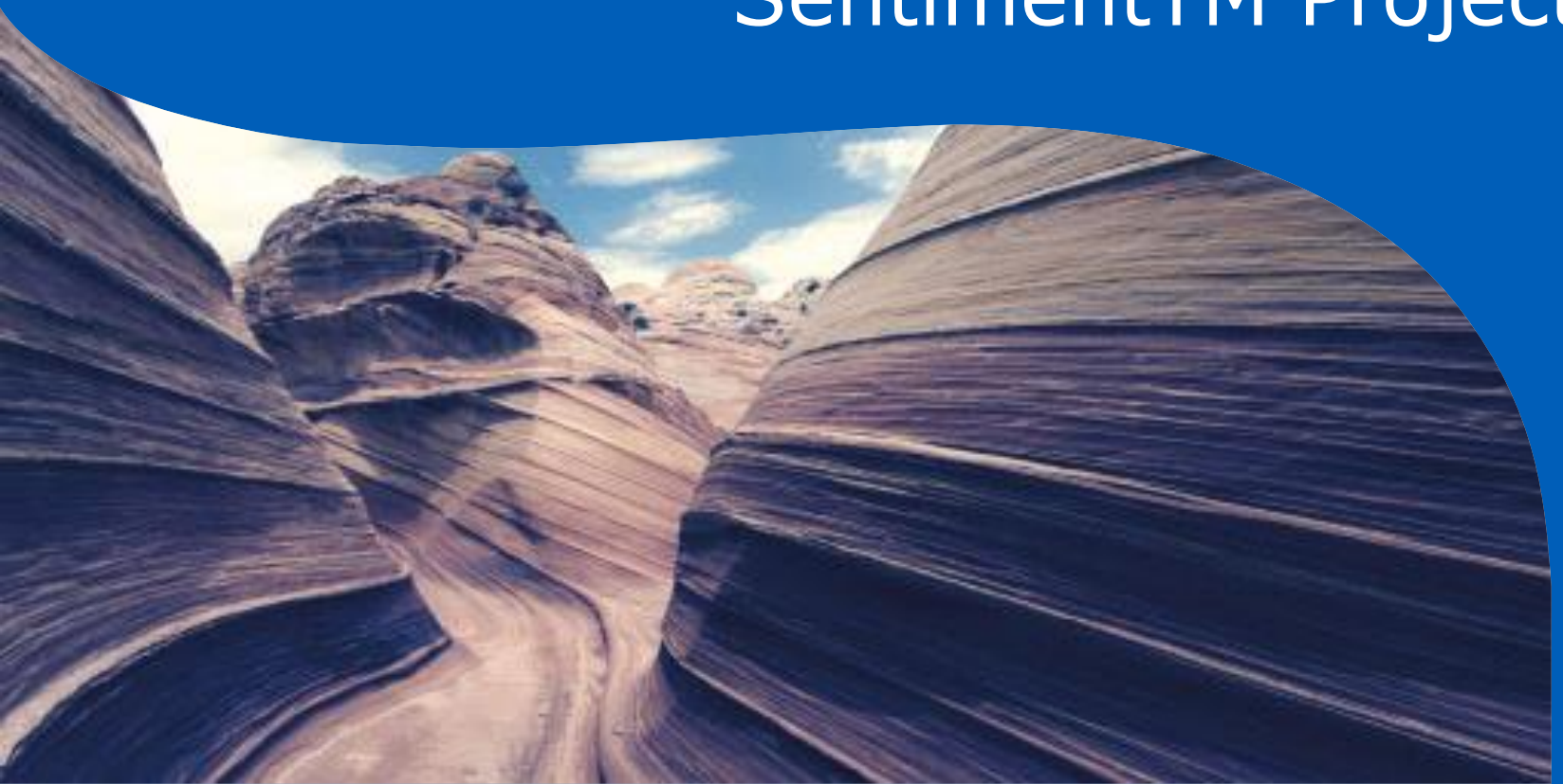
amadeus

SentimentTM Project

Sentiment Analysis

Twitter Travel Destination



The Team

Minh Phan
Intern / DTI

Alejandro Mottini
Data Scientist / R&D

Eoin Thomas
Data Scientist / DTI

Agenda

amadeus

1 Project Introduction

amadeus

Project Introduction

Travelers' sentiment?

- My direct to London Bridge is cancelled... going via Blackfriars ... if you don't see me by 2:40 I got lost :/
- Think the smell of microwaved fish is awful? The bus I'm on smells like that, first thing in the morning... #puke #tfl #London
- London st pancras train station is the most confusing place ever. Got lost about 4 times trying to find the way out

#London

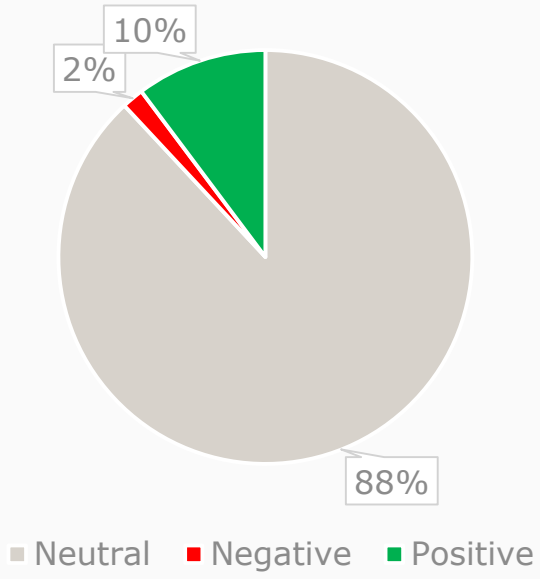
- Thing im most excited for in London? To walk across Abbey road
- Hello London !! It's been a while. Looking forward to playing at the Acklam Village Market at 5pm :) @Hot_Vox
- @LondonAsterino You are a beautiful Queen,London You sing like an angel You are awesome I will always believe in you I love you

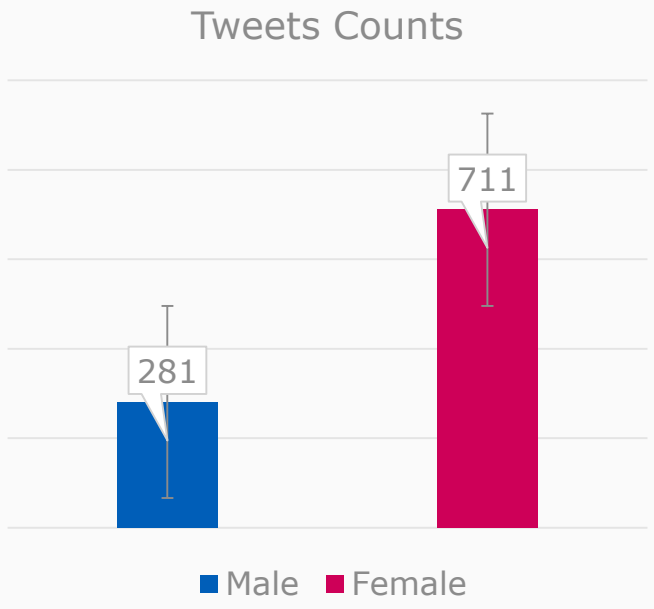
amadeus

Project Introduction

Tourism locations analysis?

Big Ben, London



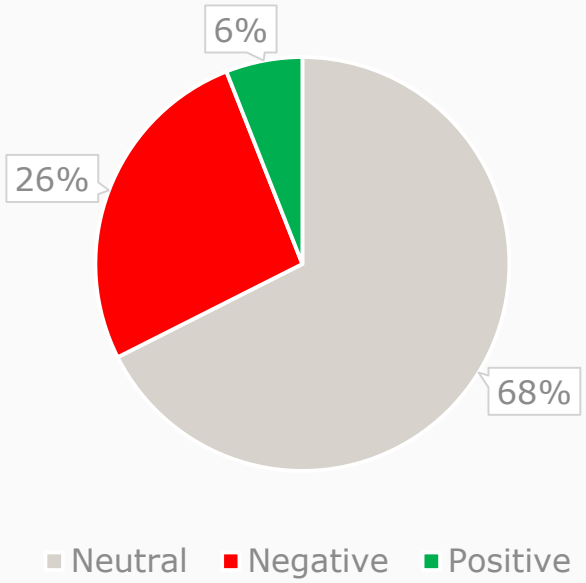


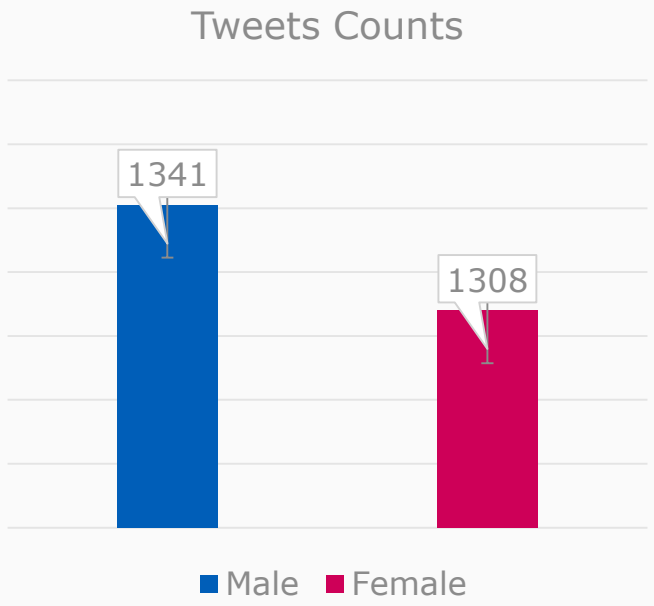
amadeus

Project Introduction

Tourism locations analysis?

Royal Borough of Greenwich, London





amadeus

Project Introduction

Targets of the project

- Explore Twitter data (API, topics, reliability…)
- Predict and classify Age, Gender and Sentiment (feeling) of travelers specific to a destination based on the information extracted from Twitter
- Build a Business Intelligence prototype (dashboard) to monitor Twitter activity related to specific destination.

amadeus

2 Twitter API, Tweets Data

amadeus

Twitter API, Tweets Data

Twitter API and Python package tweepy

- **Twitter API**
  - Free developer API vs. paid enterprise API (GNIP, https://gnip.com/)
  - REST API vs. Streaming API
  - REST API rate limit (Ref: https://dev.twitter.com/rest/public/rate-limits)
    - GET search/tweets: 450 requests / 15-min
    - GET users/lookup: 300 requests / 15-min
    - GET statuses/user_timeline: 1500 request / 15-min (max 3,200 tweets)

    - 200k tweets / 8 hours
    - 7-10 days in the past (not officially confirmed by Twitter)

- **Python package: tweepy v3.5.0**
  - Twitter account → Twitter app + key → Authorization → Query
  - Automatically pause and wait when reaching request limit
  - Support REST and Streaming APIs, return JSON format

amadeus

Twitter API, Tweets Data

Twitter data

Data including

- Tweet (text, 140 characters)
- Profile picture (image)
- Some user information (text, location, self-description, etc.)
- Account information (number of followers, following, etc.)

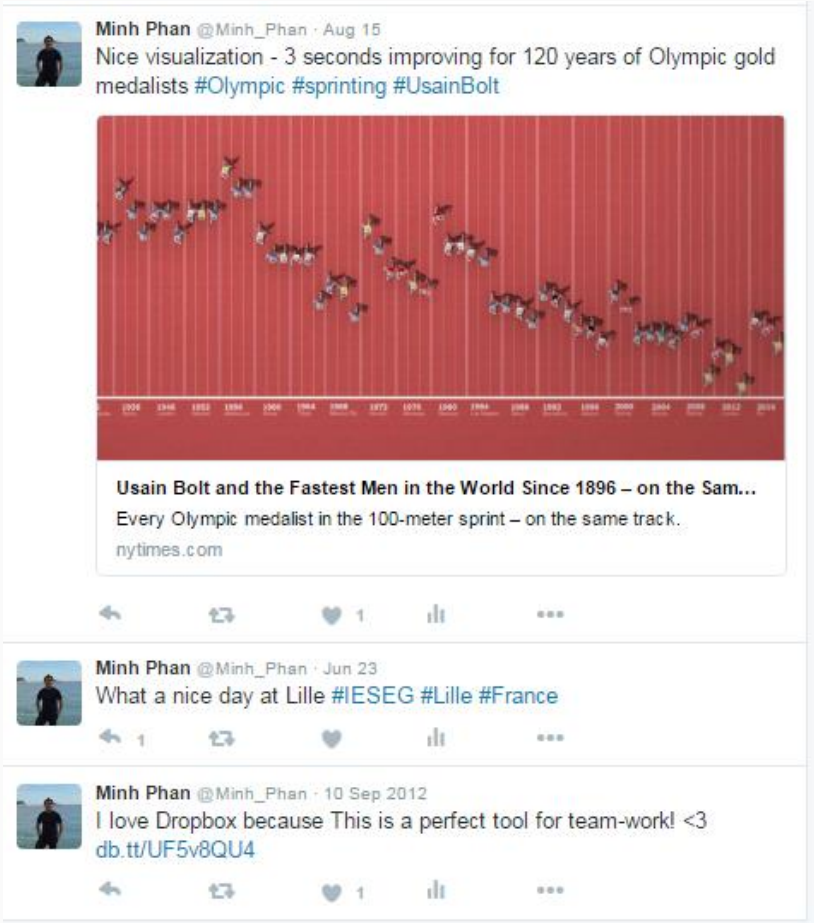
Pros

- Tons of free data
- Can access other accounts, no need to be their friends

Cons

- No gender, no birthday, no country
- Location is free text, some are unreal
- Spamming accounts

Enterprise API



amadeus

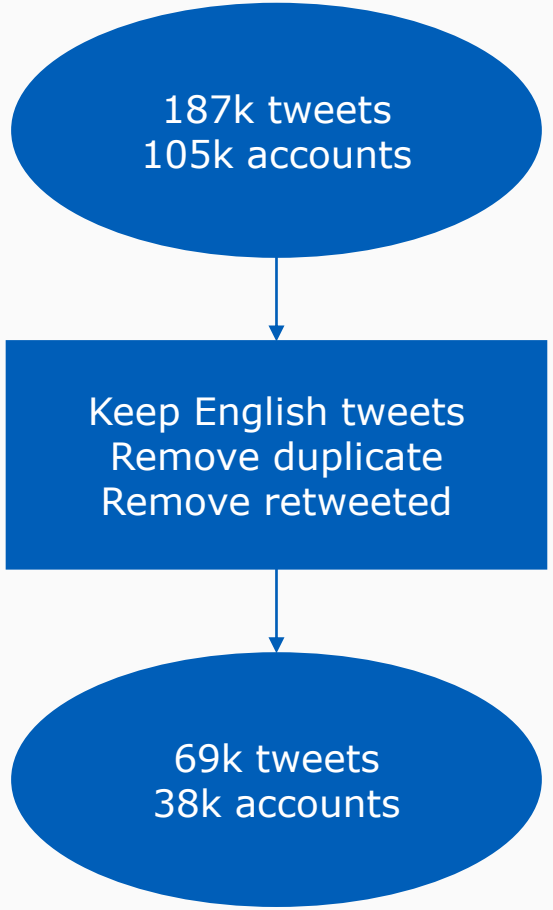
Twitter API, Tweets Data

Twitter data, some statistics

- **Tweets related to London (187k)**
  - Tweet language: en (94%), es (0.4%), de (0.4%), fr (0.3%)
  - Retweeted: 53%
  - Geo-location information: 3.1%
  - Useful tweets: 37% (69k)

- **Accounts (105k)**
  - Account language: en (84%), es (4%), de (1%), fr (1.6%)
  - Having user description: 12%
  - User profile image changed: 98%
  - Useful accounts: 36% (38k)

→ **Losing 63% of tweets**



amadeus

Twitter API, Tweets Data

Twitter data, topics clustering

Vectorize methods

- TF-IDF
- Word2Vec (Google word2vec, Stanford GloVe)

Clustering methods

- K-means
- LDA

Evaluating methods

- Inertia (sum of distances)
- Silhouette score
- Gap statistic

→ Clusters are not stable

amadeus

3 Project Pipeline

amadeus

Project Pipeline

Select a city and go with the pipeline

Collecting Data

1-2 days

tripadvisor

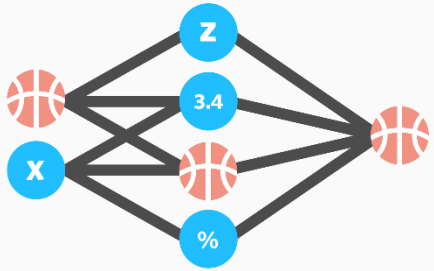
vs

Virgin atlantic

Individual vs Organisation

Classifying Models

1-2 hours



Gender, Age, Sentiment

Visualizing

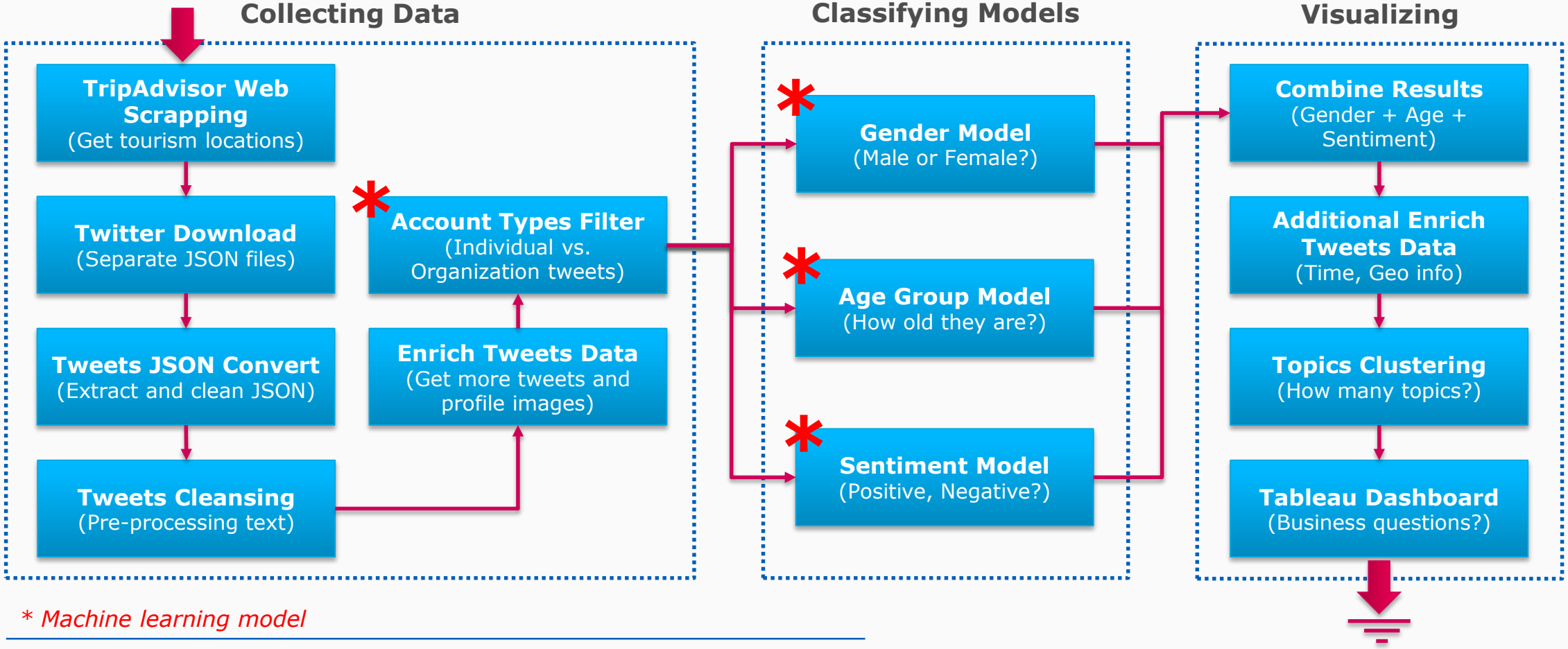
1-2 hours

+ tableau SOFTWARE

amadeus

Project Pipeline

Select a city and go with the pipeline



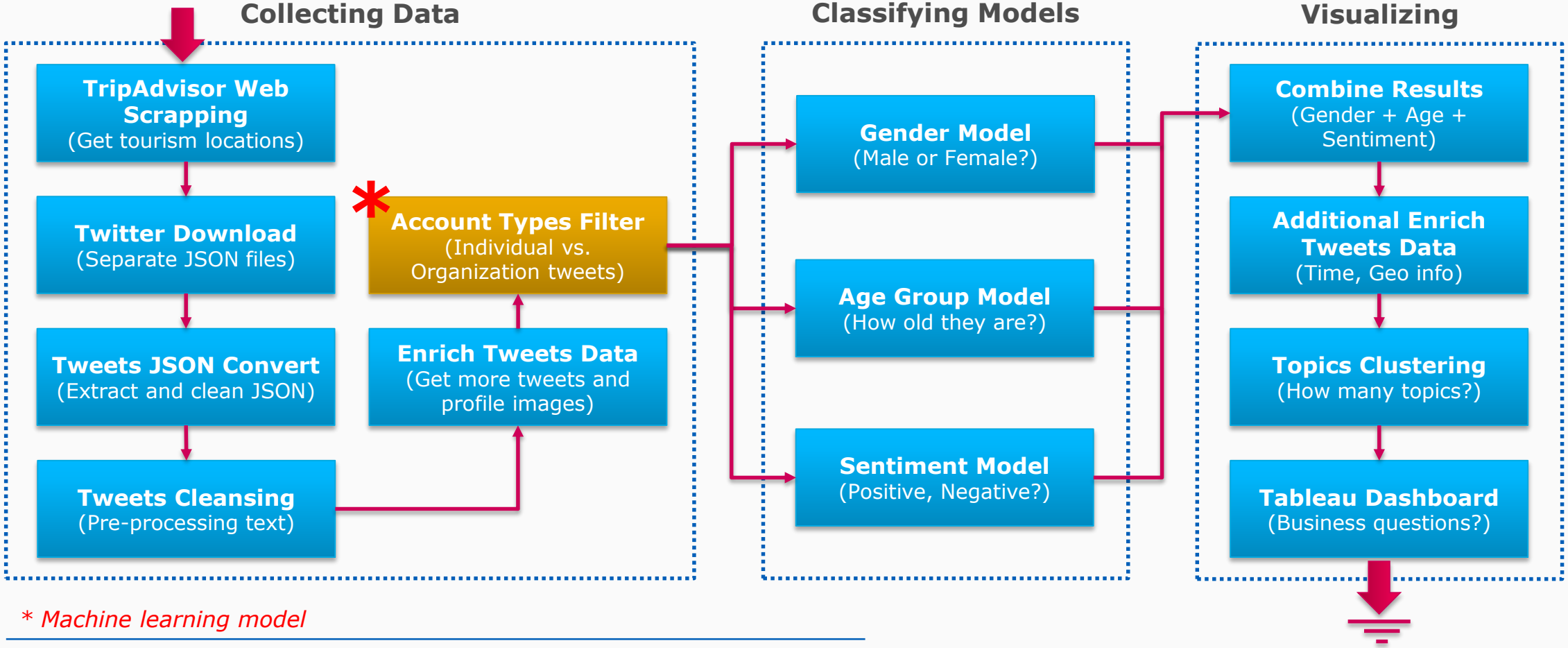
* *Machine learning model*

amadeus

4 Classification Models

Project Pipeline

Select a city and go with the pipeline



** Machine learning model*

amadeus

Classification Models

Model 1: Account types classification model

- **Training data**
  - Uniform sampling Twitter accounts, hand labelling 2,783 accounts (i.e. individual, organization, unknown)

- **Features**
  - Image features: (43 features)
    - Texture
    - Luminance (contrast)
    - HSV (hue, saturation, brightness, pleasure, arousal, dominance
    - Itten's features
  - Text features: (vector 300 dimension)
    - Google Word2Vec

Virgin atlantic	Name: VirginAtlantic / Virgin Atlantic / 20626359 Location: Worldwide Description: Hello Gorgeous! Follow us for news, banter & assistance 24/7. Visit our blog at https://t.co/ziOX3LRnnS or for official concerns visit https://t.co/ZQJmKUNzyR Tweet: A3: Secret Food Tours are a fun way to see London, just ask @cyneats @CerealKillerUK #ExpediaChat https://t.co/1Pnzy2L0tH	1.00
	Name: JimWrongUn / Jim / 769480166338854912 Location: Description: Scientist incognito. The left think I'm right; the right think I'm left. I'm stuck in the middle with you lot... Tweet: @TappnRay @LBC Oxford street is full of hare khrishnas #NottingHillCarnival #london	1.00

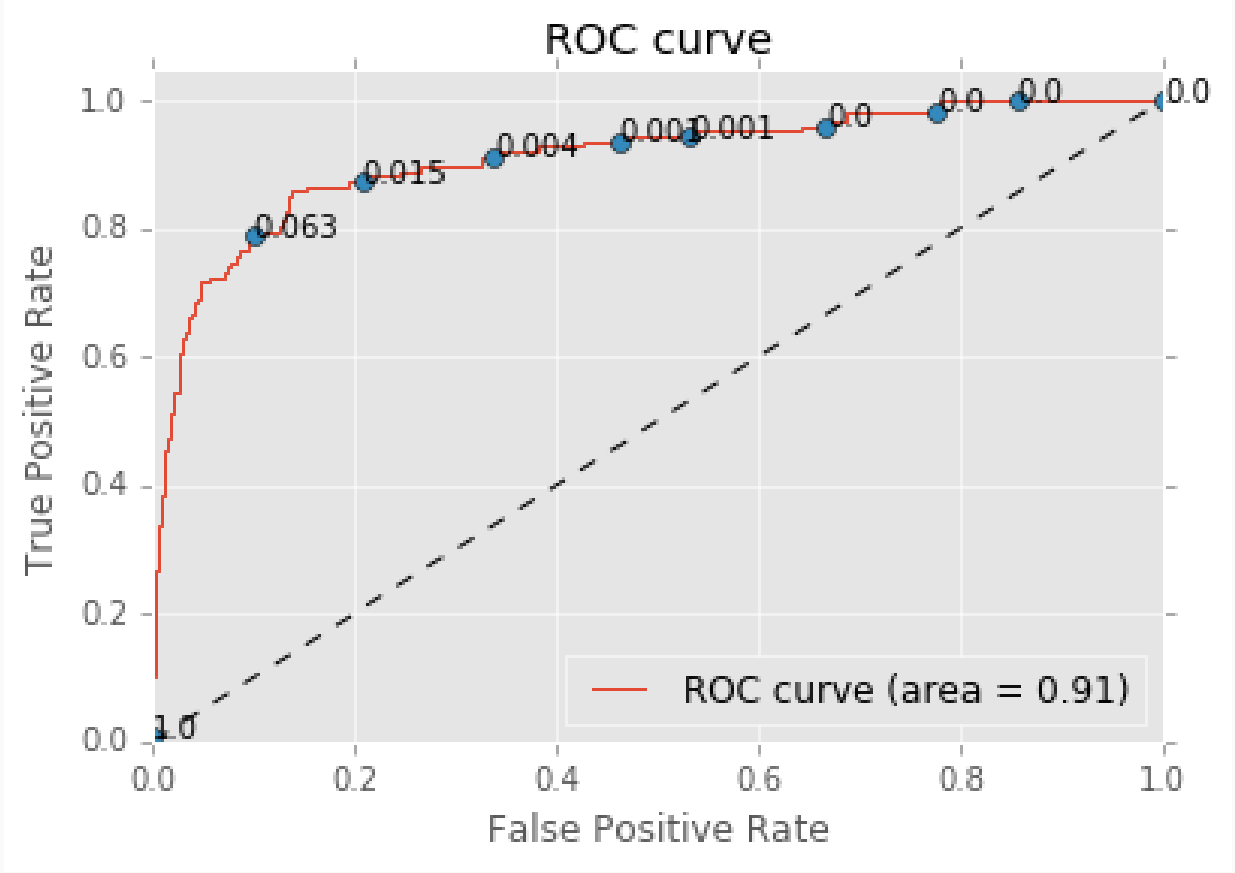
amadeus

Classification Models

Model 1: Account types classification model

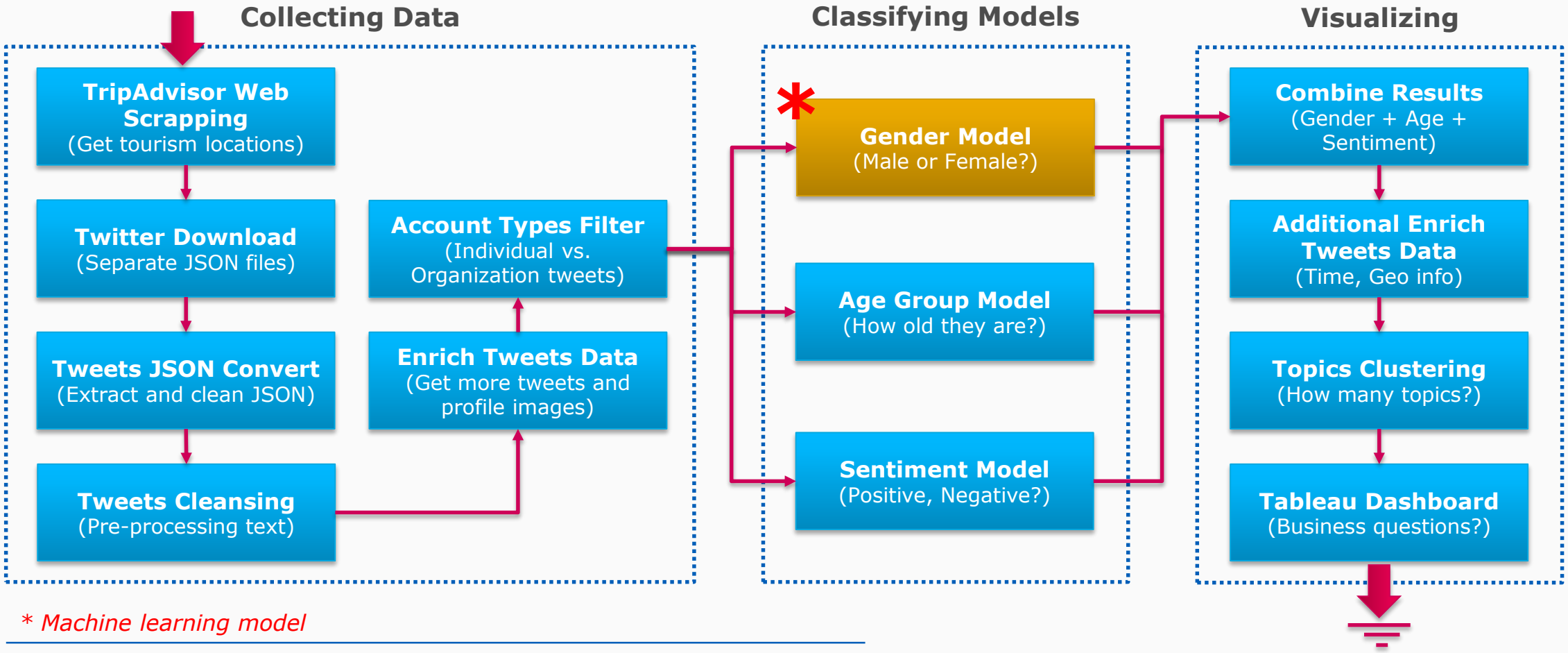
ML model
- XGBoost + GridSearchCV

Evaluation
- AUC: 0.91
- F1-score: 0.89
- Accuracy: 90%



Project Pipeline

Select a city and go with the pipeline



** Machine learning model*

amadeus

Classification Models

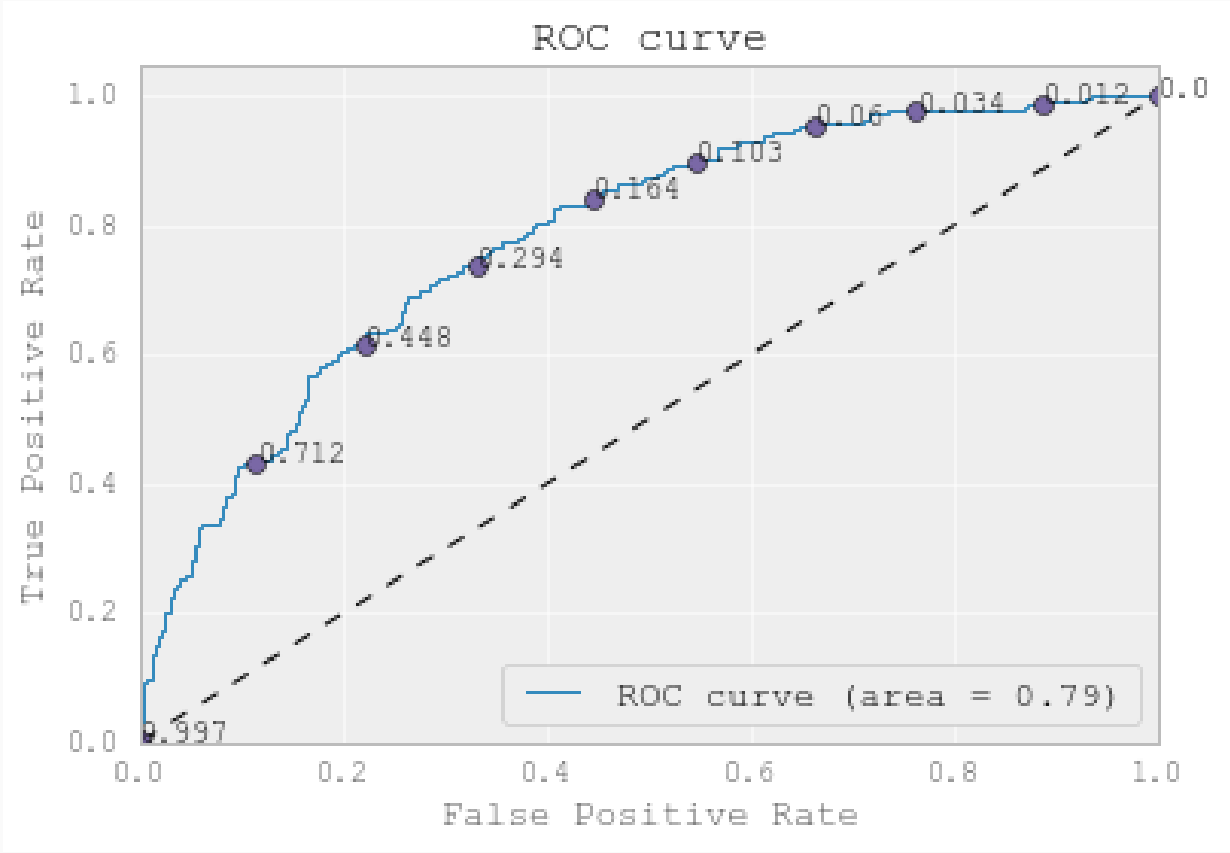
Model 2: Gender classification model

Training data
- List of Male and Female names
- Look up on Twitter: 1,901 accounts

Features
- Gender lexicon, Penn State University
- Text features:
  - Is retweeted
  - Word length
  - Count users replies, links, hashtags, emoticons
  - Count all-capital-words, e.g. AMADEUS
  - Count long-end words: e.g. greatttttt
  - Count number of Pronouns, e.g. I, he, she…

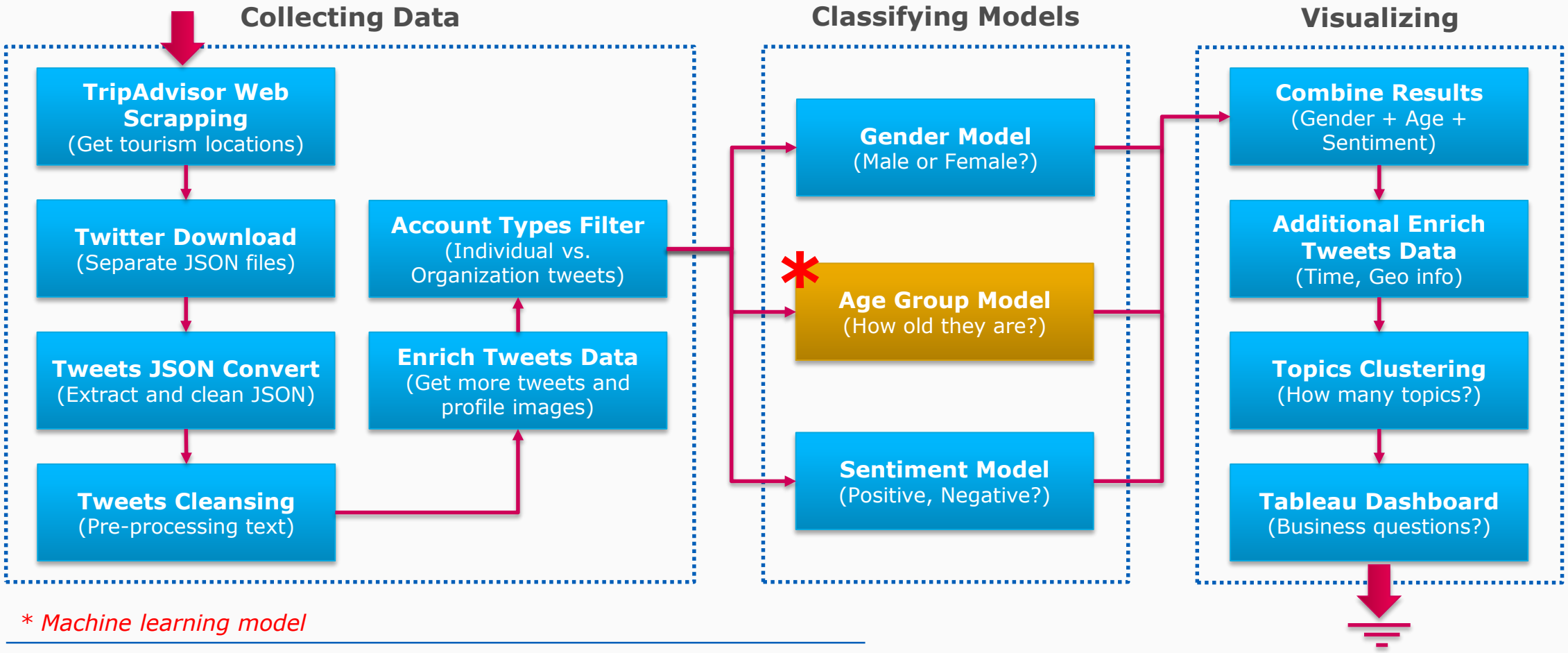
ML model
- XGBoost + GridSearchCV

Evaluation
- AUC: 0.79
- F1-score: 0.71
- Accuracy: 72%



Project Pipeline

Select a city and go with the pipeline



* *Machine learning model*

Classification Models

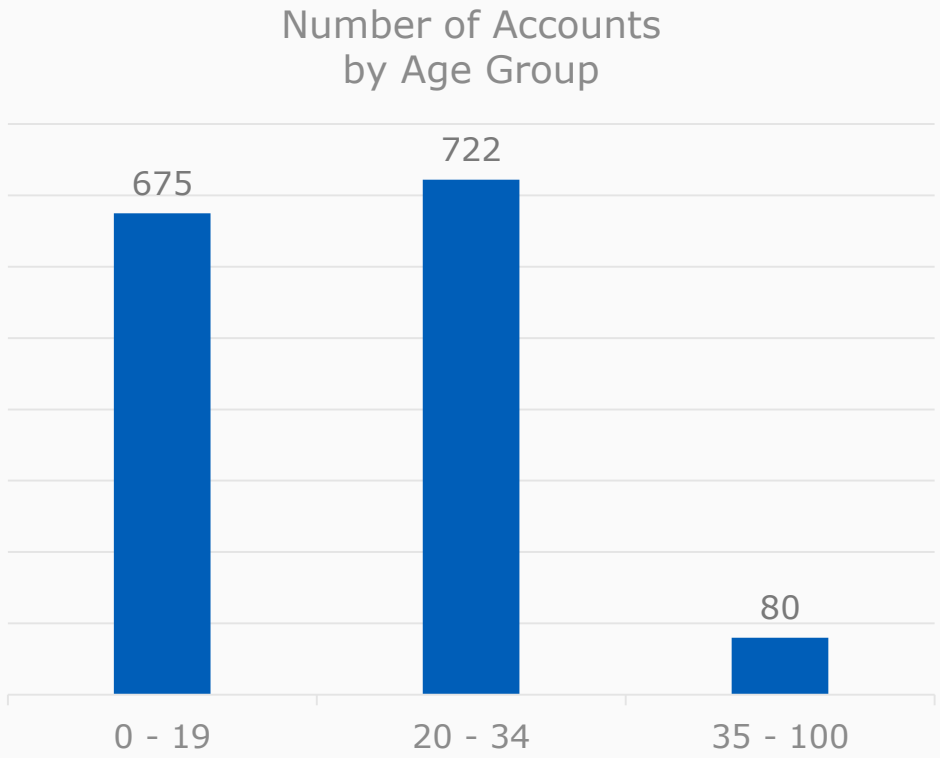
Model 3: Age group classification model

Training data
- Look up the term "Happy birthday **[age]** to me"
- 1,477 Twitter accounts

Features
- Age lexicon, Penn University
- Text features:
  - Is retweeted
  - Word length
  - Count users replies, links, hashtags, emoticons
  - Count all-capital-words, e.g. AMADEUS
  - Count long-end words: e.g. greatttttt
  - Count number of Pronouns, e.g. I, he, she…

ML model
- XGBoost + GridSearchCV

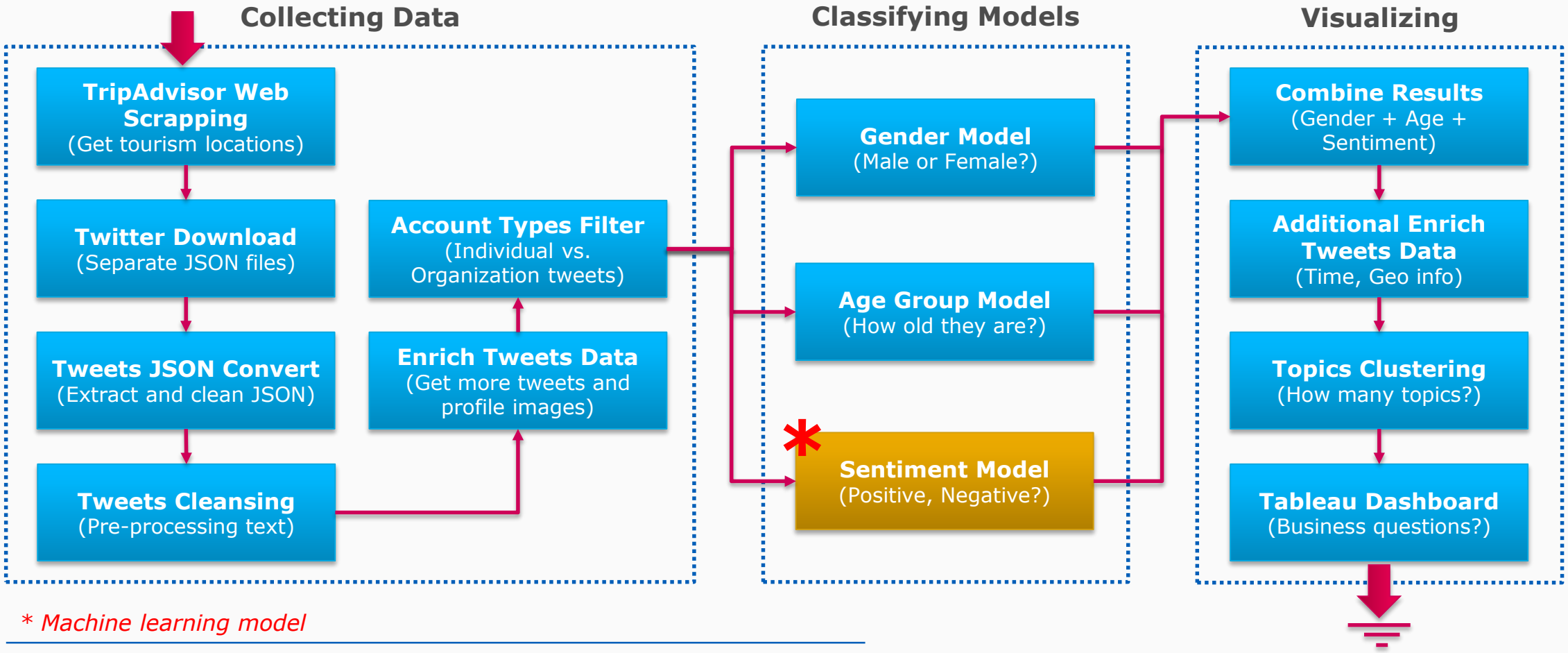
Evaluation
- F1-score: 0.63
- Accuracy: 62%



amadeus

Project Pipeline

Select a city and go with the pipeline



** Machine learning model*

amadeus

Classification Models

Model 4: Sentiment classification model

Training data

- SemEval Workshop 2013 (5,670 train, 851 test)
- SESAMm sentiment dataset (6,588 train, 1,044 test)

Features

- Lexicons:
  - NRC Hashtag Sentiment Lexicon
  - Sentiment140 Context Lexicon
  - MPQA Lexicon
  - NRC Word-Emotion Association Lexicon
  - Opinion Lexicon English
- Counts:
  - Long-end words, e.g. soooooo
  - Continuous exclamation marks, e.g. !!!???
  - Number of hashtags
  - Upper-case words, e.g. AMADEUS
- Word2Vec:
  - Twitter Word Clusters Lexicon (CMU)
  - Google Word2Vec model

Ironic or Neutral

"The view is bad but the food is great."

"Excellent! This day couldn't start off any better!"

"See London's reflection shimmers on its dark waters! Paddle down river to see the Houses of Parliament in all..."

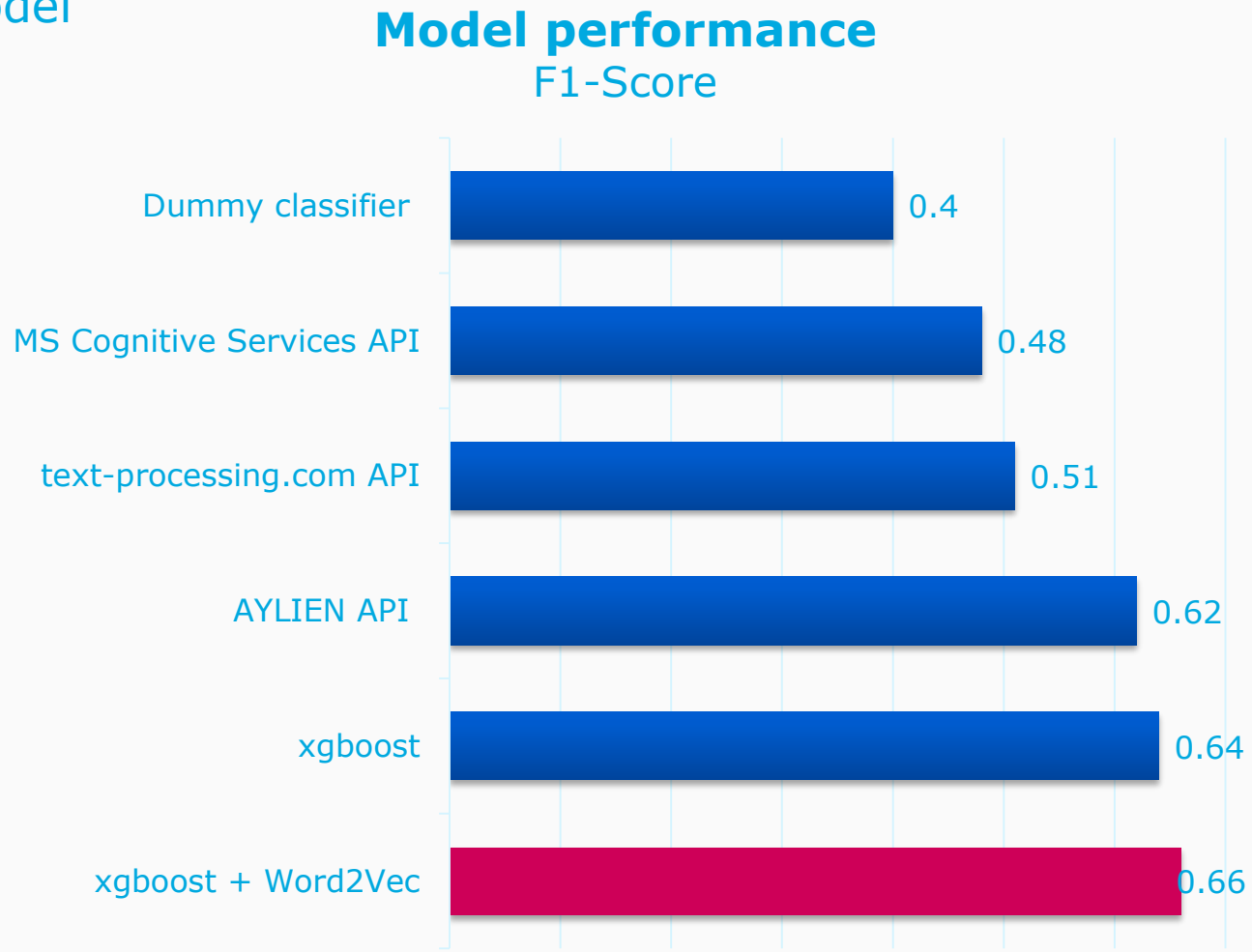
amadeus

Classification Models

Model 4: Sentiment classification model

ML model
- XGBoost + GridSearchCV

Evaluation
- F1-score 0.66
- Accuracy: 67%



amadeus

5 Product Prototype

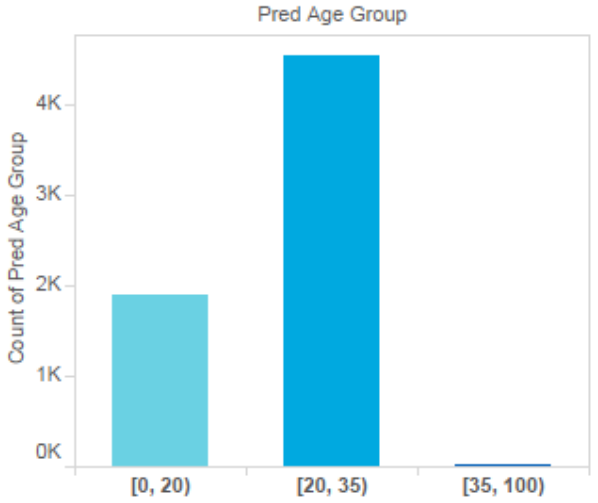
amadeus

Twitter Sentiment Analysis Project

Twitter Data Overview | Sentiment in Detail | Topic Clustering

Sentiment by Location

Age Group



Sentiment Pred

Sentiment Pred

- [ ] (All)
- [x] negative
- [ ] neutral
- [x] positive

Sentiment Pred Confidence

0.6410 — 0.9980

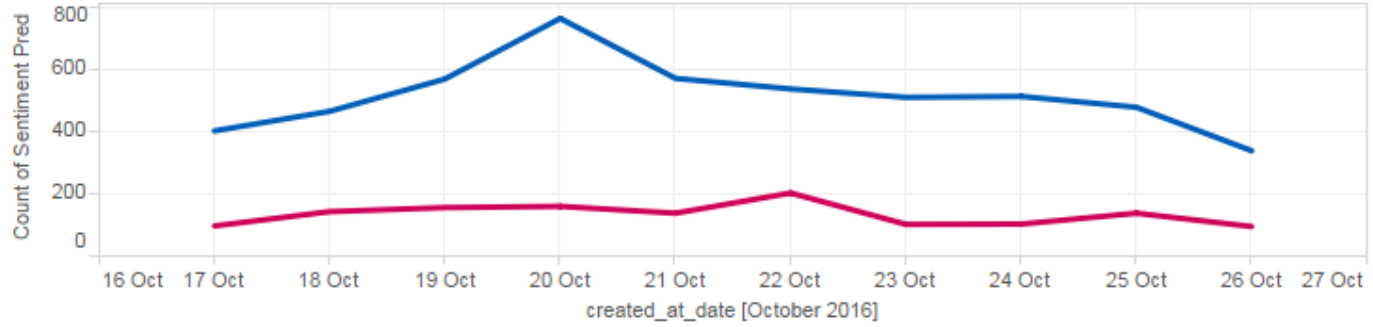
Pred Age Group

- [0, 20)
- [20, 35)
- [35, 100)

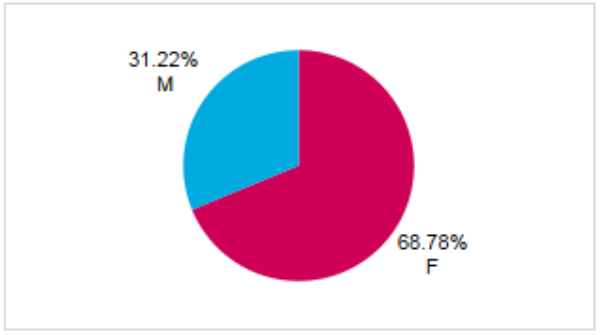
Pred Gender

- F
- M

Sentiment by Day



Gender

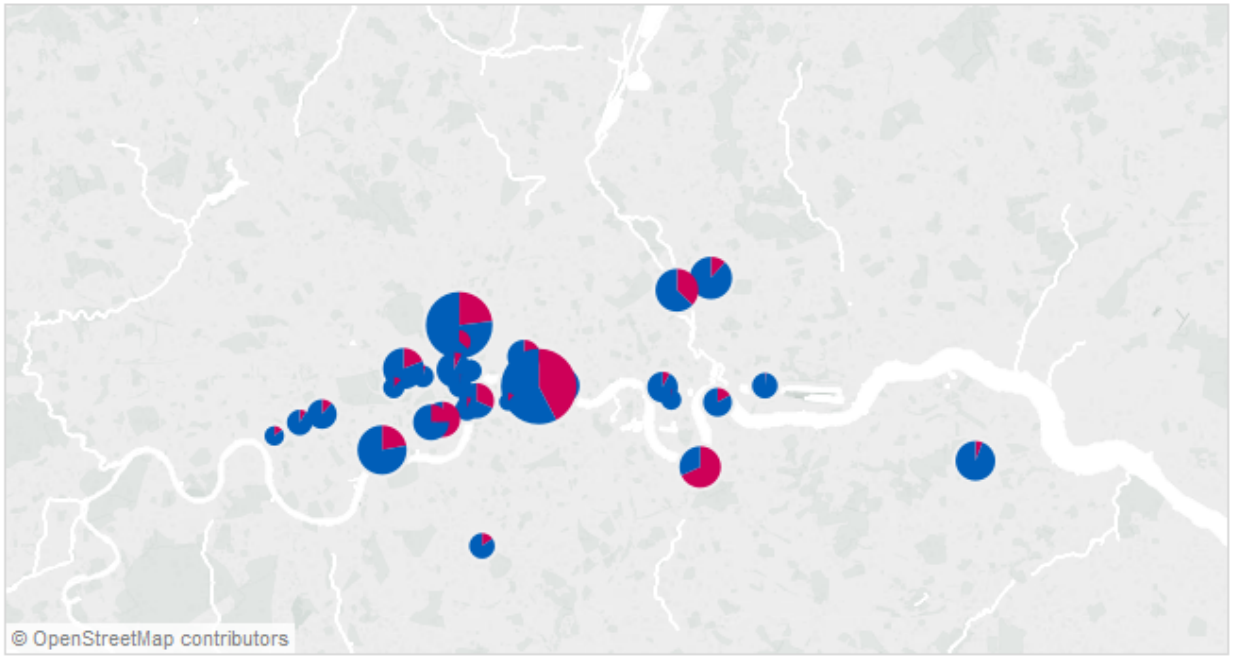


Word Counts | WordCloud Cluster 1 | WordCloud Cluster 6 | WordCloud Cluster 5 | WordCloud Cluster 9 | Dashboard 1 - Overview | Dashboard 2 - Details | Dashboard 3 - Topic | Final Story

Twitter Sentiment Analysis Project

Twitter Data Overview | Sentiment in Detail | Topic Clustering

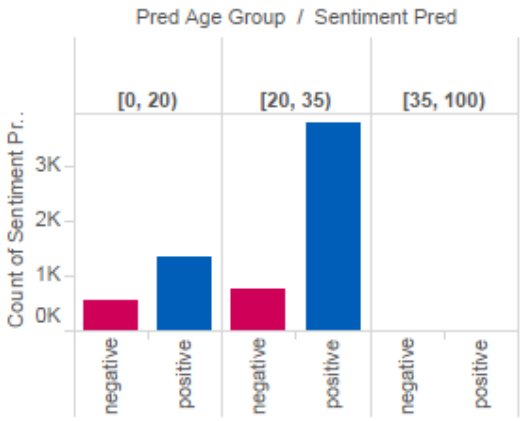
Sentiment Pie Chart by Location



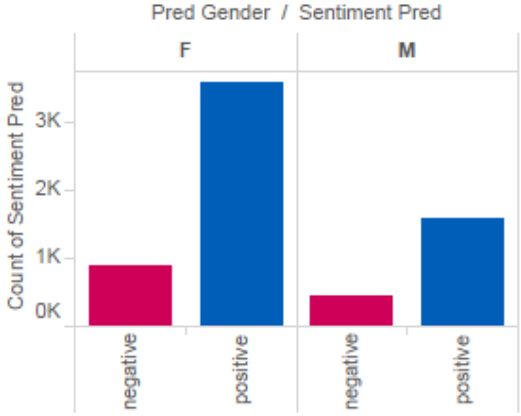
Tweets Table

User Name	
3NovicesUK	
5 live Travel	M4 : Remains closed London bound between J3, Hayes and J1, Chiswick Roundabout, queues b..
5,199-117-2,037	I liked a @YouTube video from @reversalyoutube https://t.co/49m0rlYnyT POKEMON GO IN LO..
ºGabbyº□	I wish I went to the Nickelback concert tonight in London at the o2 cry

Age Group vs. Sentiment



Gender vs. Sentiment



Sentiment Pred

- negative
- positive

Sentiment Pred

- (All)
- negative
- neutral
- positive

Sentiment Pred Confidence

0.6410 — 0.9980

Geopy Address

- (All)
- Null
- Abbey Road, Roding Riverside, Thames View, Lo...
- Albert Bridge, Chelsea, Royal Borough of Kensing...
- Almeida Theatre, Almeida Street, Mildmay Park, I...
- Angel, Islington High Street, Pentonville, London ...
- Apple Market, Covent Garden Piazza, St Clement ...
- ArcelorMittal Orbit, 3, Thornton Street, Stratford M...
- Arcola Theatre, 24, Ashwin Street, South Hornsey...
- Arctic Spas, 20, Kingsway, City Trading Estate, H...
- Arts Theatre, 6-7, Great Newport Street, St Cleme...
- Baker Street, Marylebone, City of Westminster, Lo...
- Bank of England Museum, Bank, Bartholomew La...

Limit

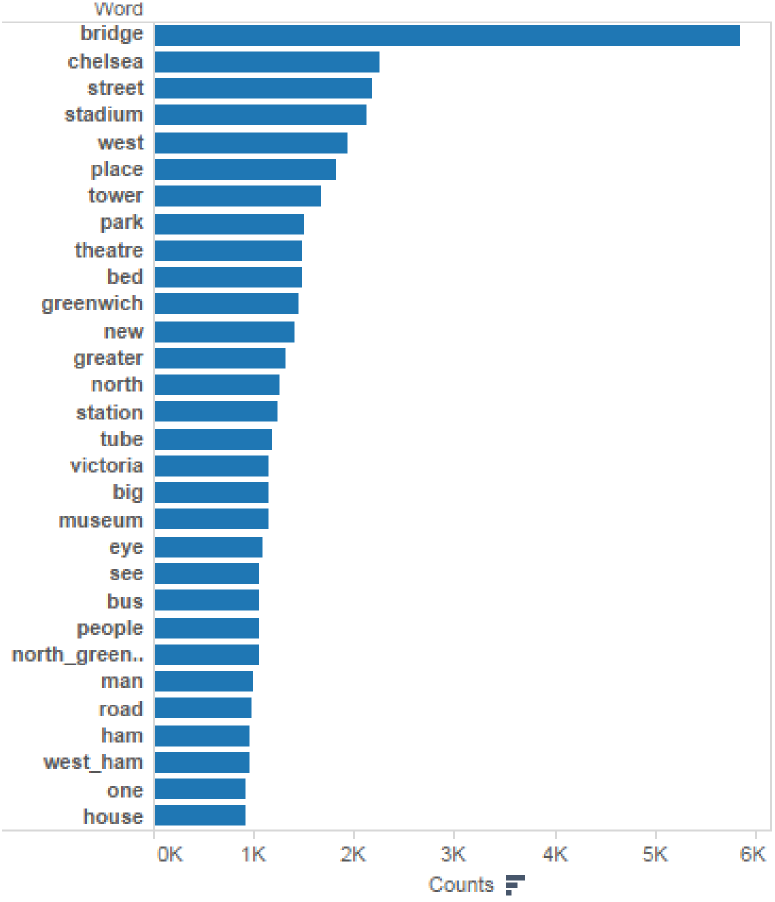
Top 30 by SUM([Number of Records])

Sentiment Pie Chart by Location | Tweets Table | Gender vs. Sentiment | Age Group vs. Sentiment | Word Counts | WordCloud Cluster 1 | WordCloud Cluster 6 | WordCloud Cluster 5 | WordCloud Cluster 9

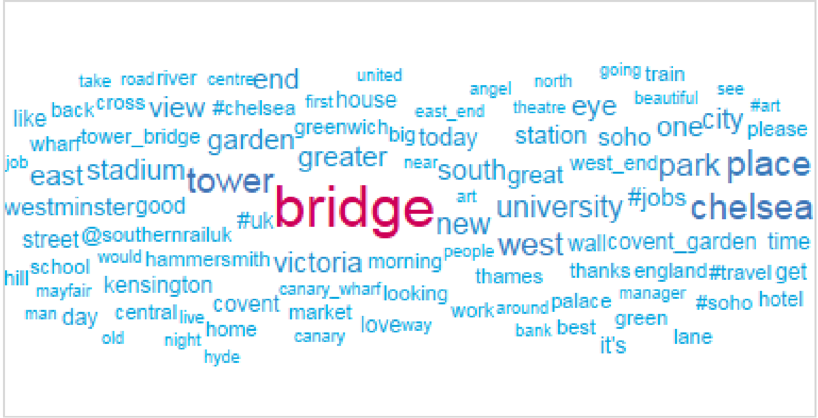
Twitter Sentiment Analysis Project

Twitter Data Overview | Sentiment in Detail | Topic Clustering

Word Counts

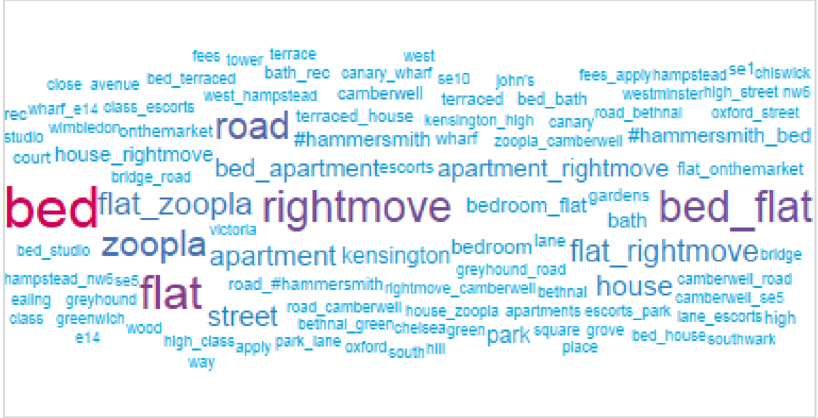


WordCloud Cluster 1

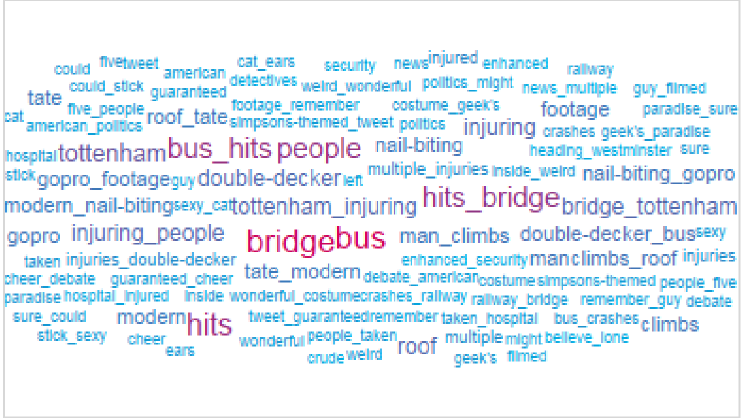


WordCloud Cluster 6

WordCloud Cluster 5



WordCloud Cluster 9



Sentiment Pie Chart by Location | Tweets Table | Gender vs. Sentiment | Age Group vs. Sentiment | Word Counts | WordCloud Cluster 1 | WordCloud Cluster 6 | WordCloud Cluster 5 | WordCloud Cluster 9

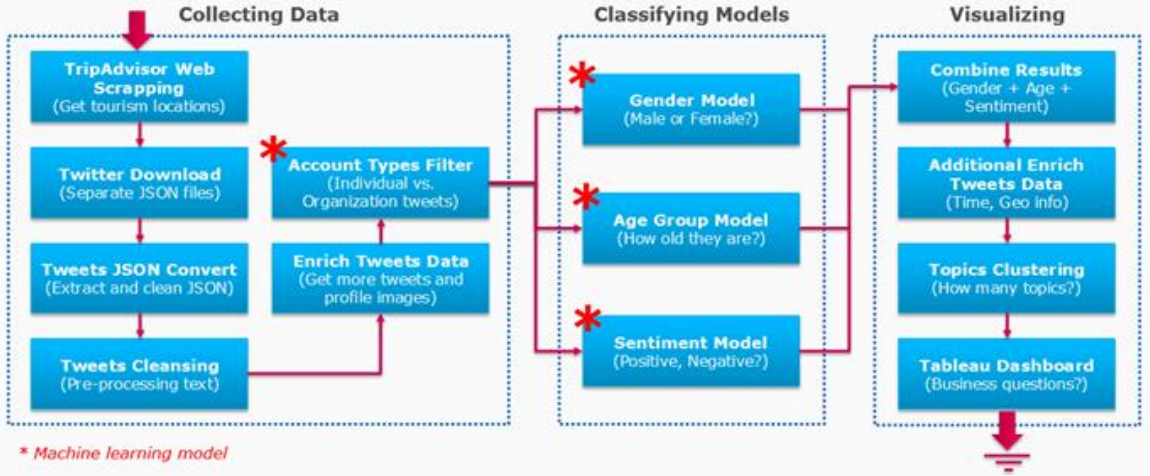
6 Wrap-up

amadeus

Wrap-up

Things we did

- Built a complete pipeline for the project, easy to add or improve modules in the future
- Built an Account Types classification model, AUC=0.91, Accuracy=**90%**
- Built a Gender classification model, AUC=0.79, Accuracy=**72%**
- Built a Age Group classification model, F1-Score=0.63, Accuracy=**62%** (21% better than a dummy model)
- Built a Sentiment Classification model, F1-Score=0.66, Accuracy=**67%**, better than commercial APIs (e.g. Microsoft Cognitive API, AYLIEN API), free API (e.g. text-processing.com API)



amadeus

Wrap-up

Things to improve

- More data to improve the 4 models
  - More ways to collect data, e.g. “London” :)
  - Buy data from Twitter enterprise API (GNIP)
  - Other training data sources, e.g. Sentiment140 http://help.sentiment140.com/for-students/

- Upgrade Machine Learning model
  - Deep learning model to improve Age and Gender classification
  - Ensemble 16 models can increase Sentiment classification model’s F1-Score up to 0.70

- Add company/user history database to save time and avoid duplicated work

- Tweak decision threshold to fit with the local data distribution, e.g. percentage of organization vs. individual accounts is different by city

amadeus

_____ #ThankYou! and #MerryChristmas!

You can follow us on:
AmadeusITGroup

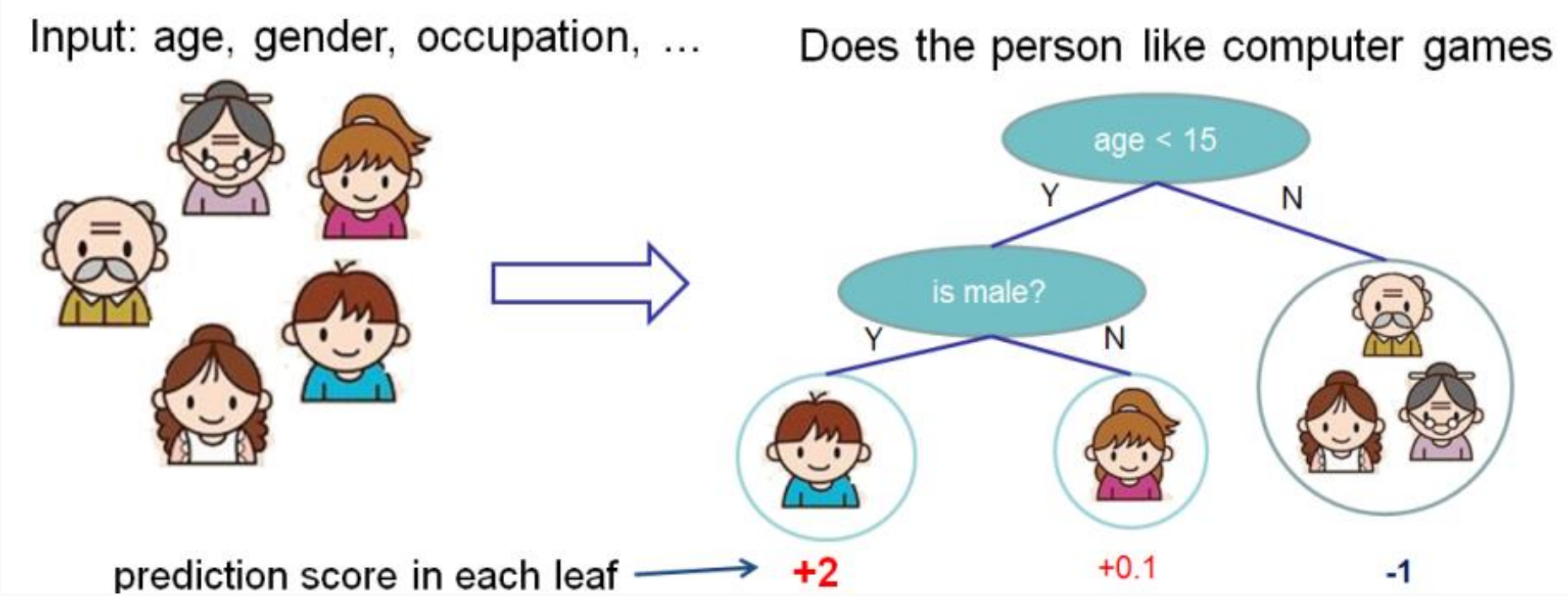
amadeus.com/blog
amadeus.com

amadeus

Appendix 1

XGBoost - Extreme Gradient Boosting

Tree Ensemble

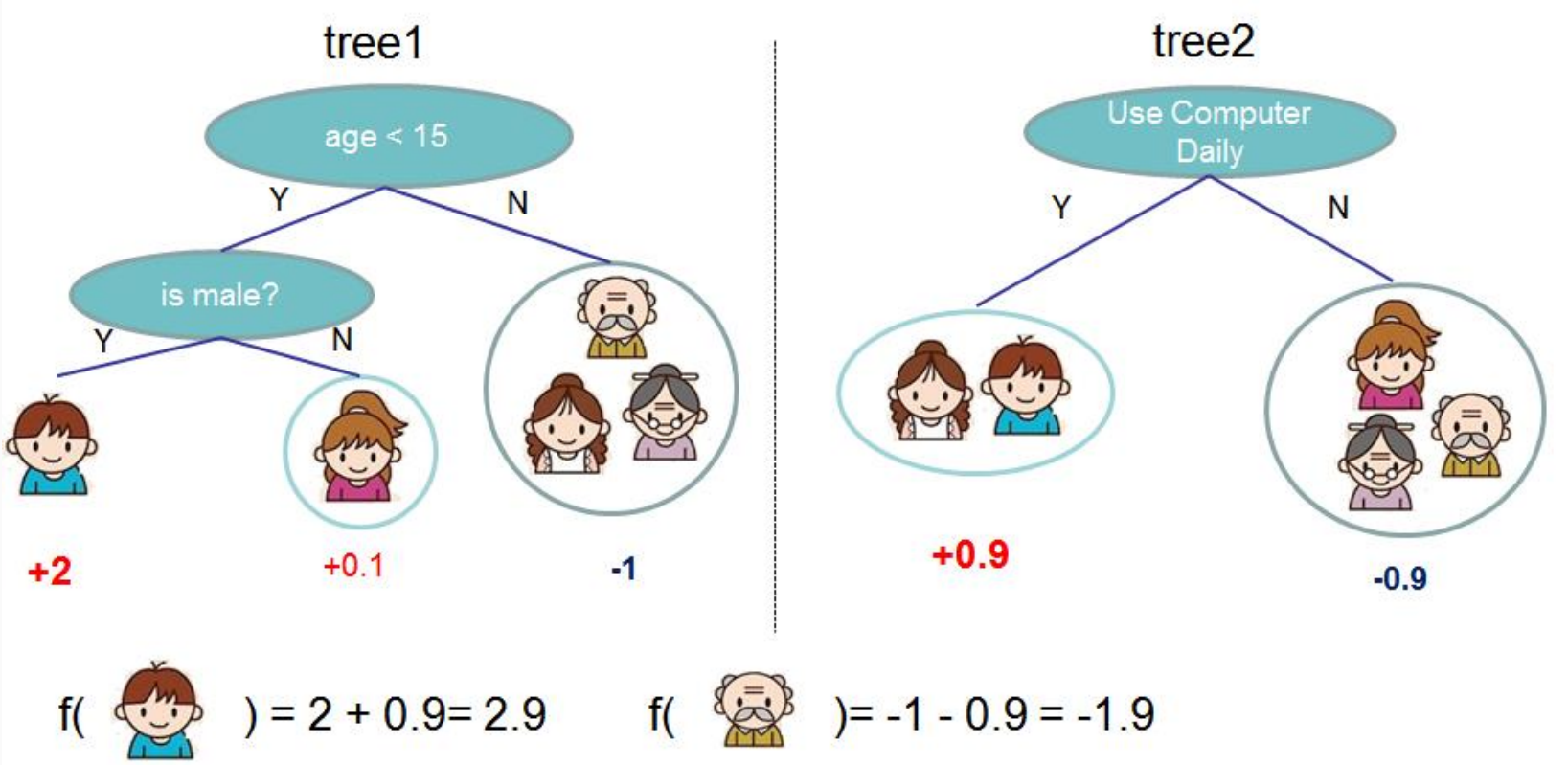


amadeus

Appendix 1

XGBoost - Extreme Gradient Boosting

Tree Ensemble



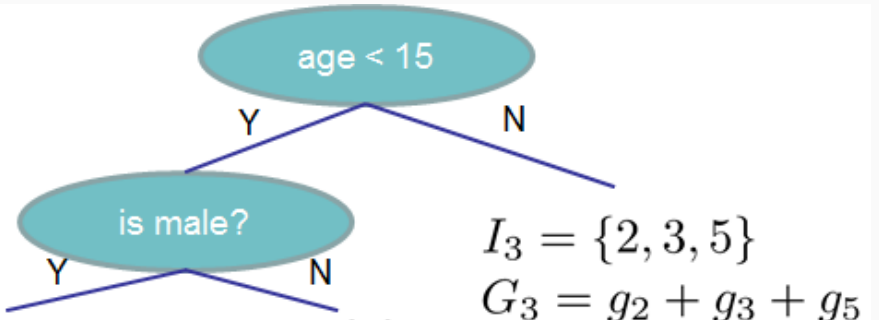
amadeus

Appendix 1

XGBoost - Extreme Gradient Boosting

Gradient Boosting

Instance index		gradient statistics
1		g1, h1
2		g2, h2
3		g3, h3
4		g4, h4
5		g5, h5



$I_1 = \{1\}$
$G_1 = g_1$
$H_1 = h_1$

$I_2 = \{4\}$
$G_2 = g_4$
$H_4 = h_4$

$I_3 = \{2, 3, 5\}$
$G_3 = g_2 + g_3 + g_5$
$H_3 = h_2 + h_3 + h_5$

$$Obj = -\sum_j \frac{G_j^2}{H_j + \lambda} + 3\gamma$$

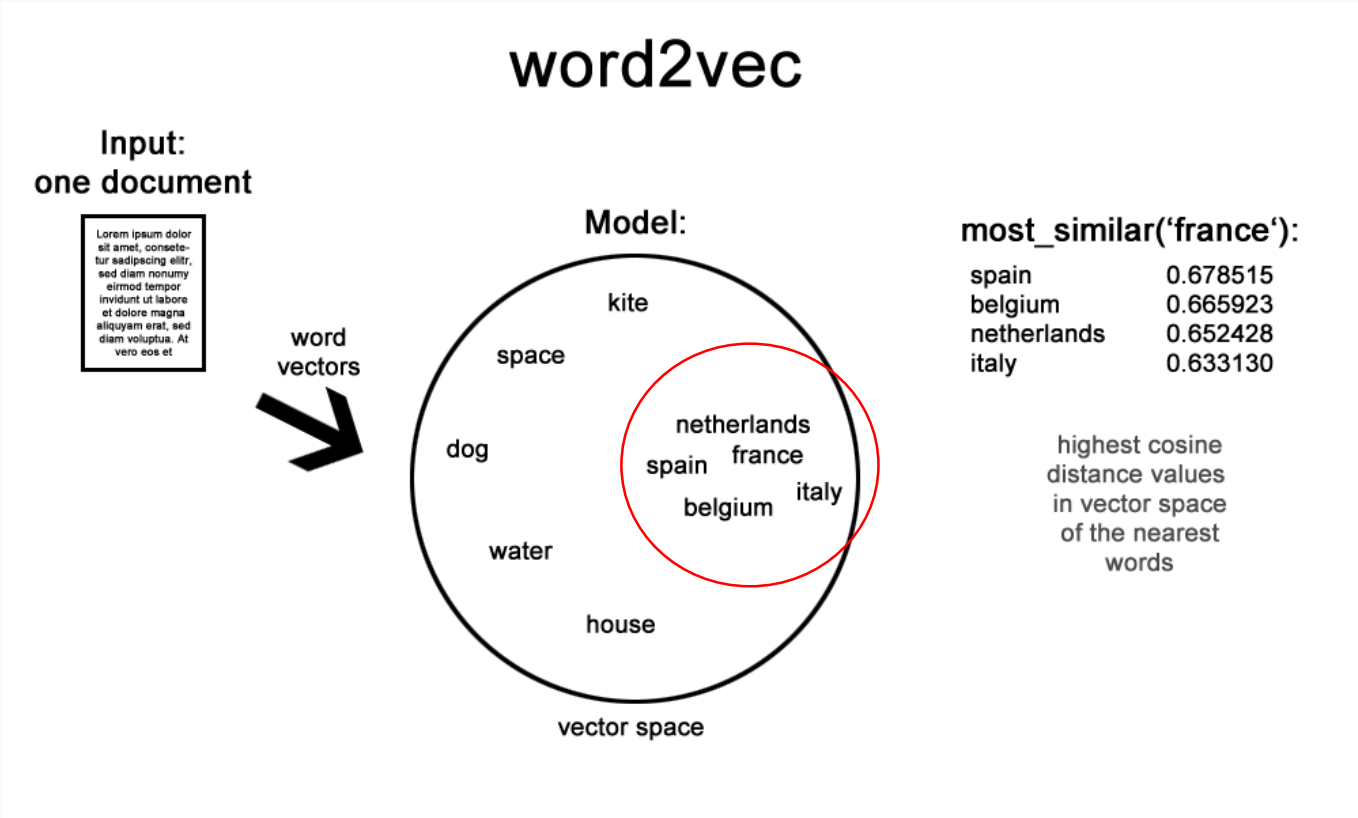
The smaller the score is, the better the structure is

amadeus

Appendix 2

Word2Vec

- Represent a word by a vector (vectorize)
- Reduce the data dimension compare with other word frequency method (TF-IDF)
- Better results most of the time



amadeus