Understanding Happiness Using NLP

By: Nadia & Manal

As Project 4 of SDAIA Data Science Bootcamp (T5)

START

01

OUTLINE:

- Introduction
- Process & Methodology
- Result
- Conclusion

Introduction

Being **happy** is one of the most fundamental requirements of a living being. Since the beginning of the human civilization, man has also been trying to develop new technologies, make new tools and improve his lifestyle for the sole purpose of attaining happiness. However, in its race of scientific endeavor and pursuit of money and luxuries, man is hardly aware of what exactly constitutes happiness .

Project Goals: Topic modelling to understand what makes Amazon's workers happy for the previous 24 hours

PROCESS & METHODOLOGY

1
Tools

2
Data

3
EDA

4
NLP

Tools

- Panda
- Numpy
- Matplotlib and Seaborn
- Gensim
- Spacy
- Nltk ' Natural Language Toolkit '
- pyLDAvis

Data

Data source: **kaggle** [https://www.kaggle.com/ritresearch/happydb]

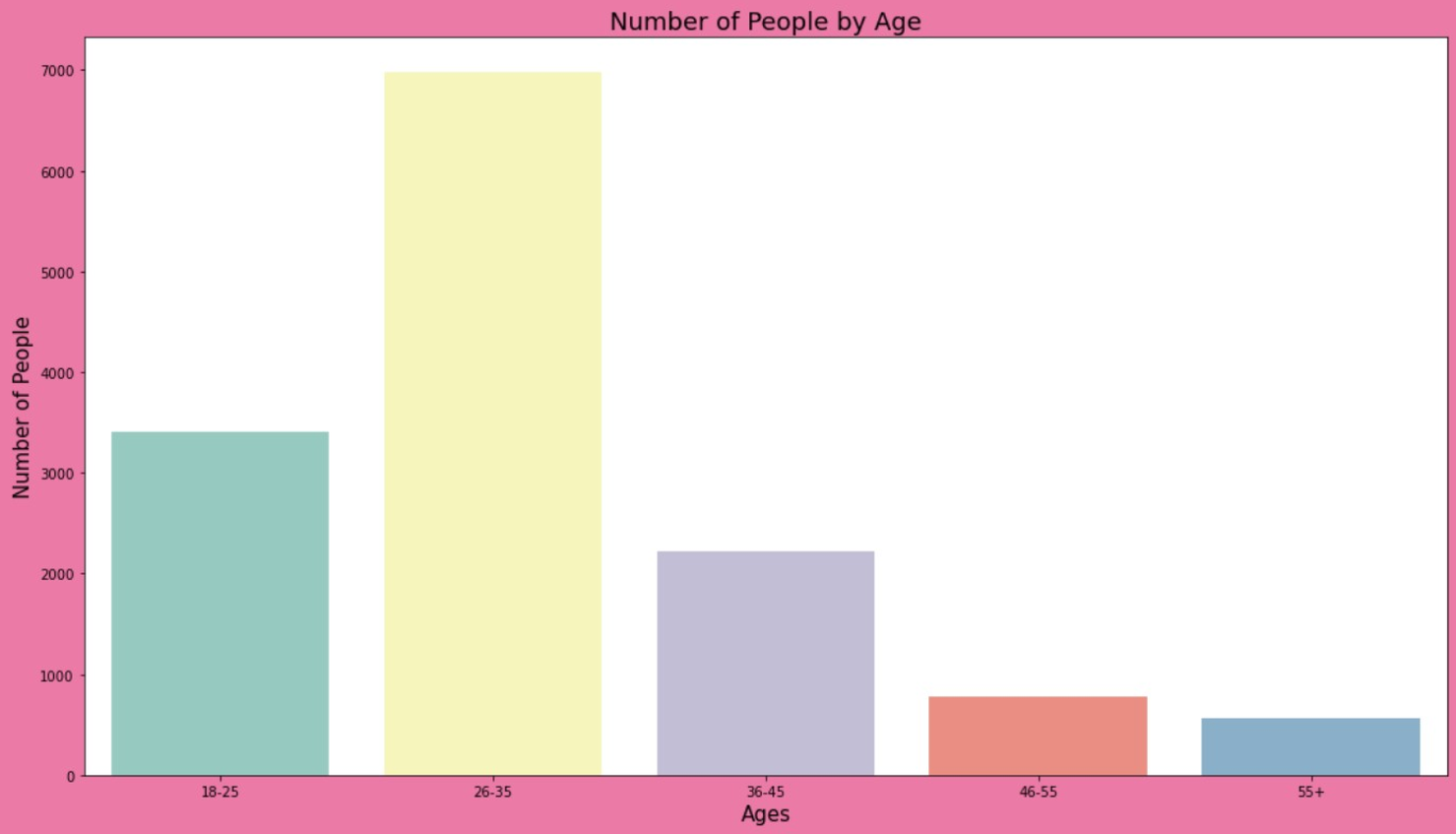
2 data files :

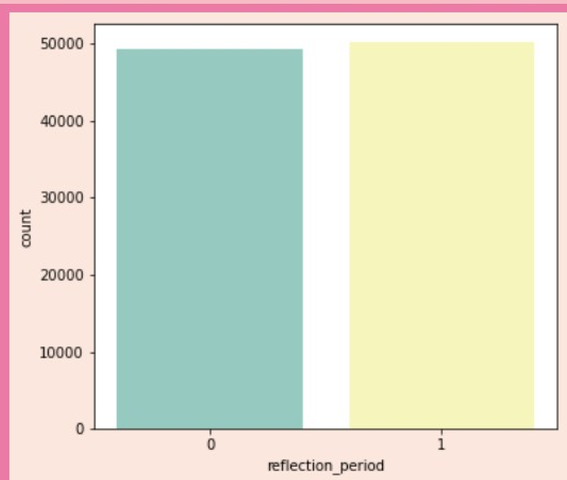
- cleaned_hm.csv: the cleaned-up corpus of 100,000 crowd-sourced happy moments.
- demographic.csv: the demographic information of the worker who provided the moment.

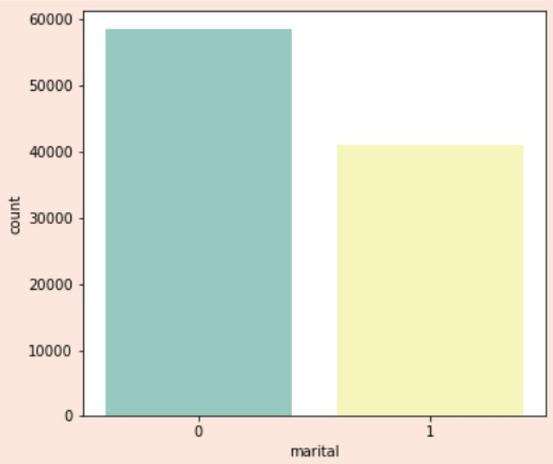
100,535 observations & 14 features

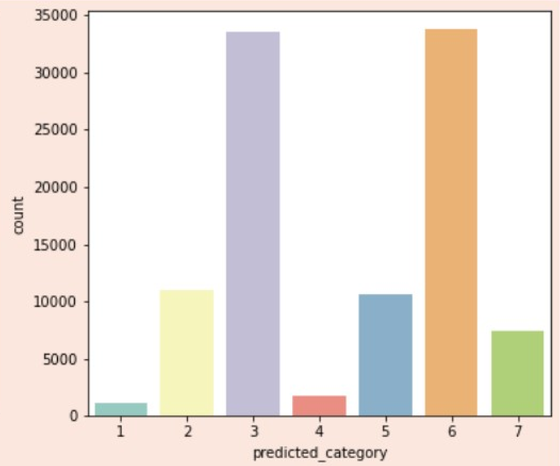
EDA

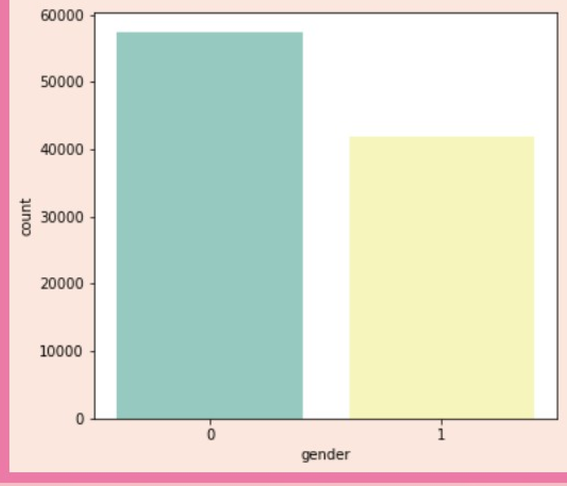
- After merging, cleaning, dropping nulls and unneeded columns and encoding categorical features, we explored the data.
- interested in:
  - number of people patricipated basec on: age, gender, country, parenthood and martial

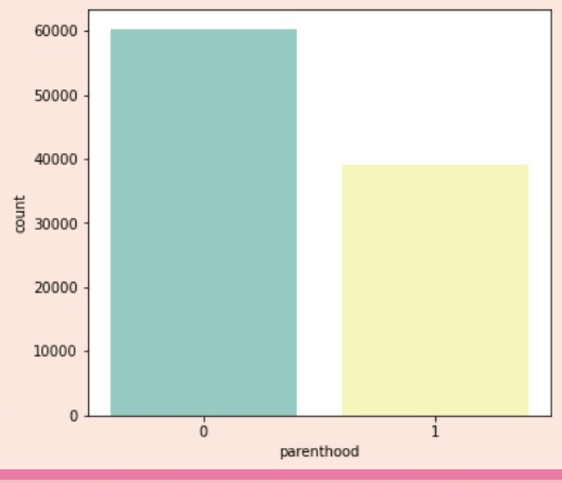

Number of People by Age
Number of People
7000
6000
5000
4000
3000
2000
1000
0
18-25
26-35
36-45
46-55
55+
Ages

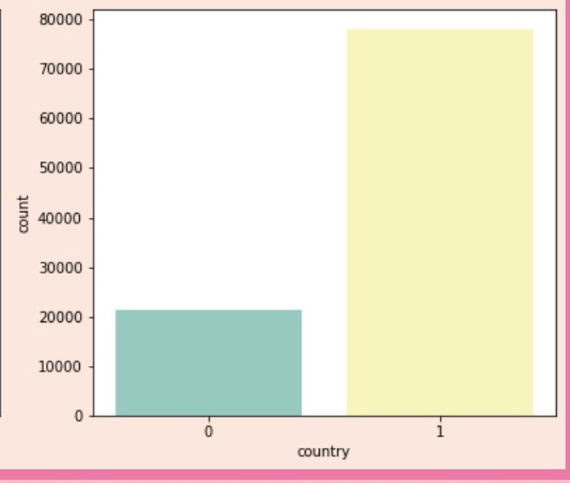

count
reflection_period
0
1
10000
20000
30000
40000
50000


count
marital
0
1
10000
20000
30000
40000
50000
60000


count
predicted_category
1
2
3
4
5
6
7
5000
10000
15000
20000
25000
30000
35000


count
gender
0
1
10000
20000
30000
40000
50000
60000


count
parenthood
0
1
10000
20000
30000
40000
50000
60000


count
country
0
1
10000
20000
30000
40000
50000
60000
70000
80000

NLP

WORKFLOW

Preprocessed Data

NLTK

Topic Model

Gensim LDA

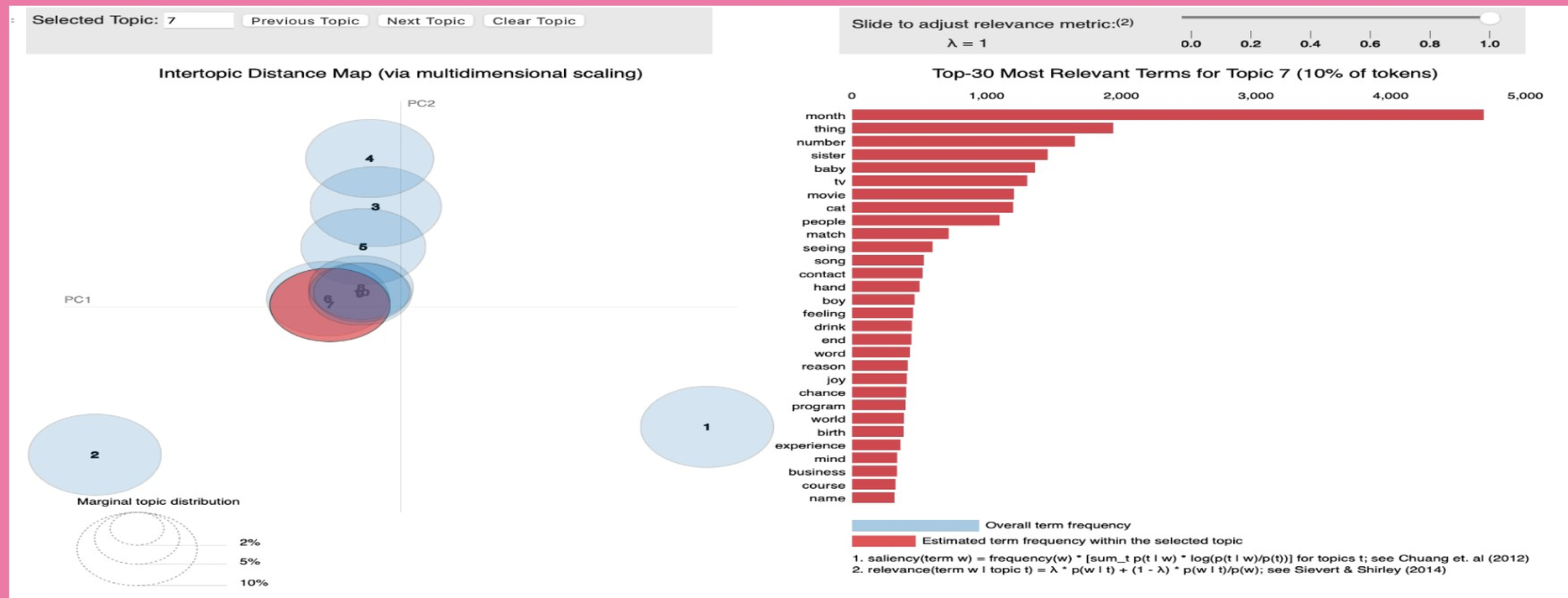
Evaluation & Visualization

WordCloud pyLDAvis
Coherence Score

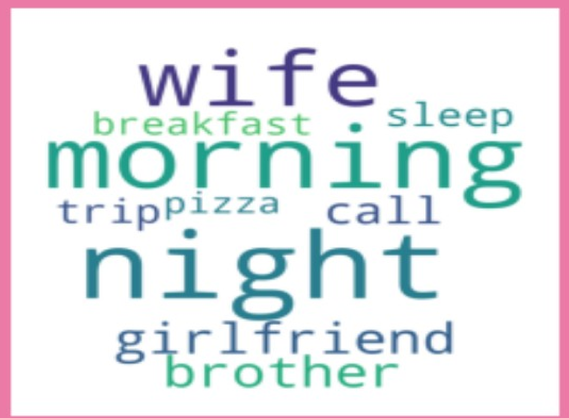
Topics in LDA model

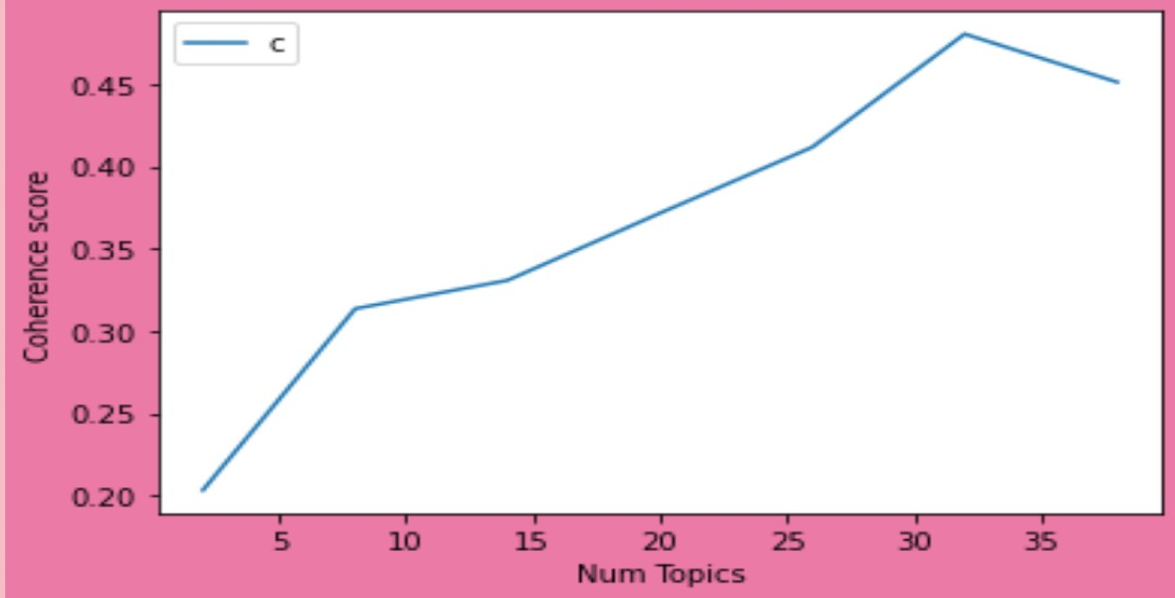
```
Topic: 0
Words: 0.172*"home" + 0.074*"money" + 0.070*"birthday" + 0.056*"house" + 0.049*"boss"
+ 0.047*"vacation" + 0.040*"girl" + 0.029*"mturk" + 0.024*"saving" + 0.022*"plan"


Topic: 1
Words: 0.146*"night" + 0.126*"morning" + 0.096*"wife" + 0.054*"girlfriend" + 0.054*"br
other" + 0.039*"call" + 0.031*"trip" + 0.030*"sleep" + 0.027*"breakfast" + 0.025*"pizz
a"


Topic: 2
Words: 0.097*"game" + 0.044*"video" + 0.036*"company" + 0.034*"project" + 0.024*"team"
+ 0.023*"help" + 0.022*"student" + 0.021*"health" + 0.020*"music" + 0.020*"need"


Topic: 3
Words: 0.196*"time" + 0.092*"daughter" + 0.070*"event" + 0.046*"lunch" + 0.043*"boyfri
end" + 0.030*"kid" + 0.030*"child" + 0.025*"show" + 0.021*"bed" + 0.020*"way"


Topic: 4
Words: 0.209*"work" + 0.127*"today" + 0.091*"job" + 0.073*"family" + 0.020*"class" + 0
.018*"gift" + 0.014*"store" + 0.014*"interview" + 0.012*"computer" + 0.012*"bonus"
```

TOPICS VISUALIZATION with pyLDAvis



WordClouds based on LDA model






```
Num Topics = 2  has Coherence Value of 0.2032
Num Topics = 8  has Coherence Value of 0.3134
Num Topics = 14  has Coherence Value of 0.3306
Num Topics = 20  has Coherence Value of 0.3714
Num Topics = 26  has Coherence Value of 0.4115
Num Topics = 32  has Coherence Value of 0.4803
Num Topics = 38  has Coherence Value of 0.4511
```

Evaluation by Coherence score

Topic classification based on LDA model + Domain Knowledge

```
# Topics Classification :
Topic 0: spending
Topic 1: relaxing
Topic 2: entertainment
Topic 3: Social Activities
Topic 4: Achievement
```

The optimal topics number here is 32 topics with coherence score = 0.48

Conclusion

Thank You !