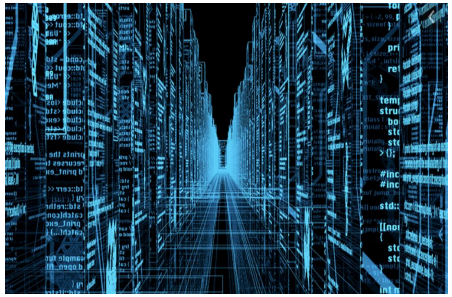
Evaluation of Company's Information Privacy Agreement

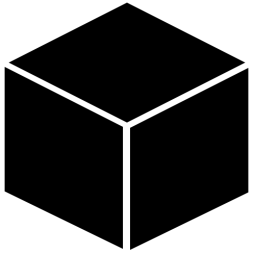
UDAP Final Project

Yuwei Zhu, Tingting Gu

Incentives



There is **large volume of data** being collected everyday from us, either through digital devices, applications or websites.



What Information is **collected** and How?

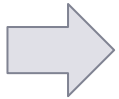
Privacy Policy Agreement is usually **hard to read** and easily neglected

“

There are certain **topics** within **companies’** information privacy agreements that are **concerned** by the customers.

Datasets and Questions

- ACL/COLING 2014
  - 1,010 privacy policies from top websites ranked on Alexa.com
- Twitter API Data
  - Extraction of tweets corresponding to topics identified from policy corpus

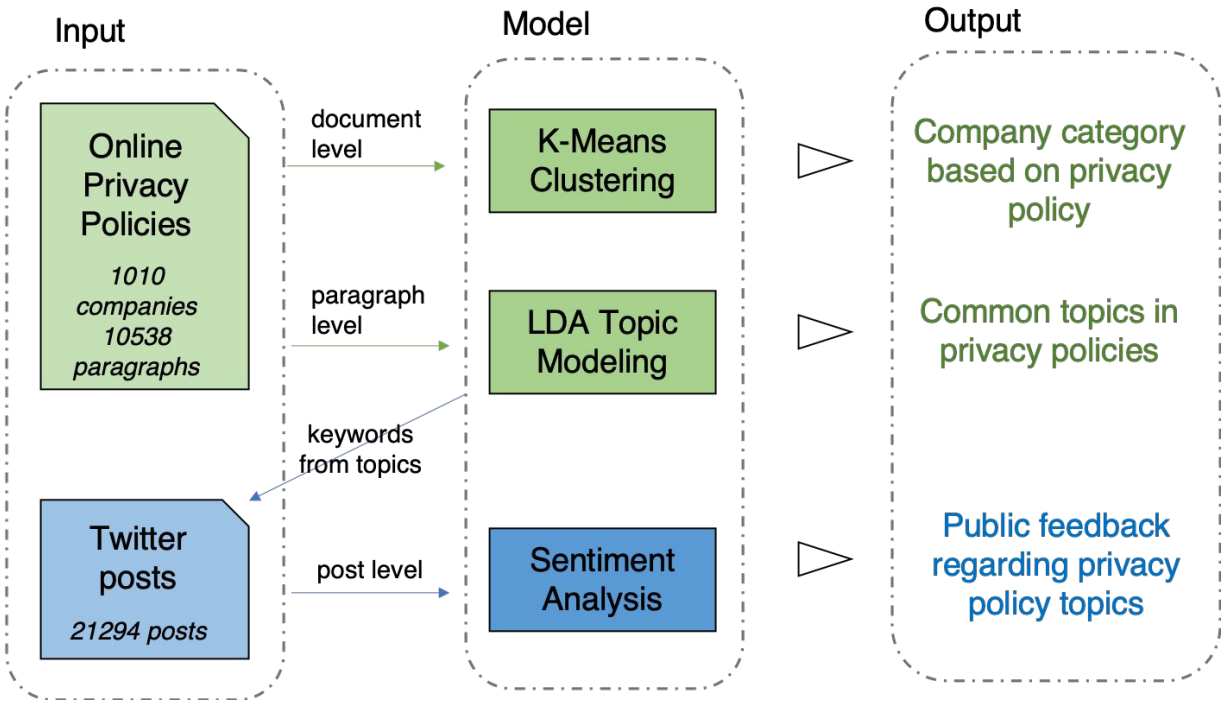


Company Cluster

Topics from privacy documents

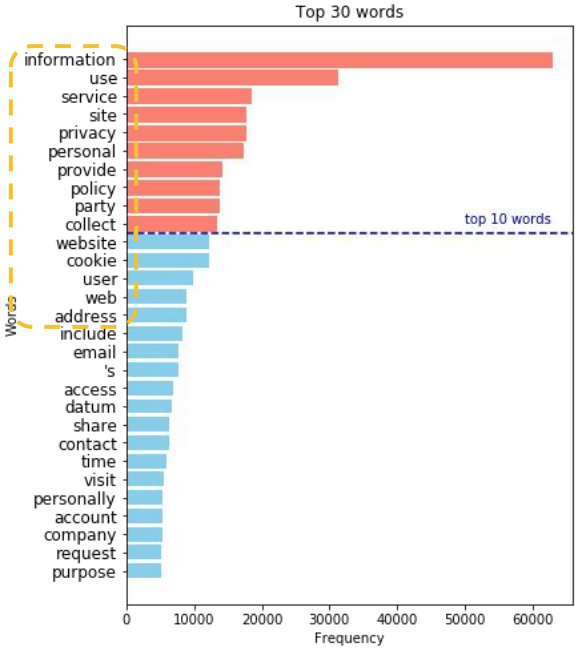
Sentiment on Topics

Analysis Pipeline

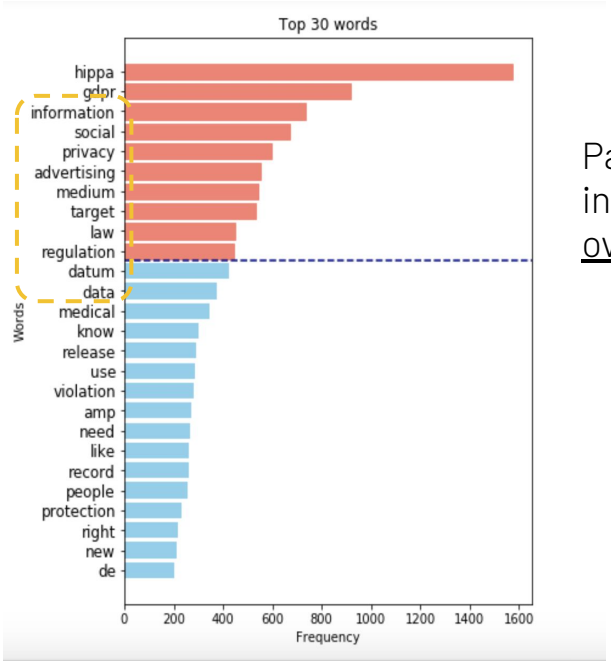


Basic Text Analysis

- ACL/COLING 2014



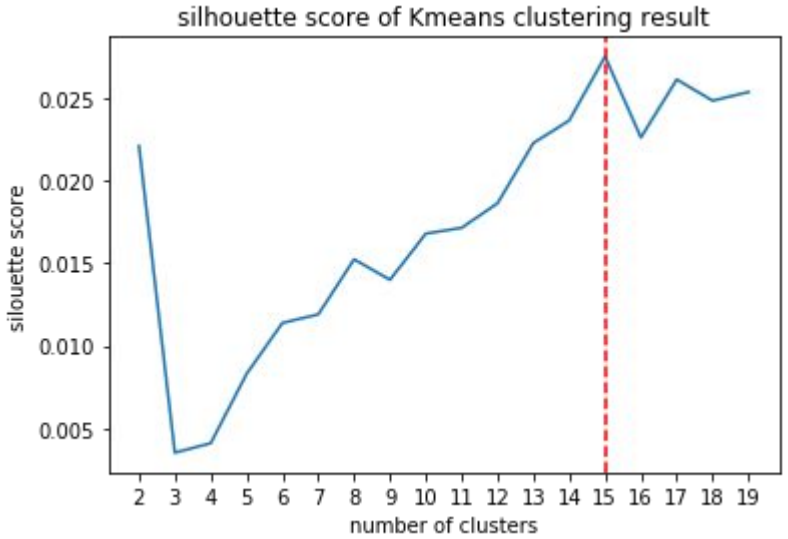
- Twitter API Data



Part of top words in two datasets overlapped.

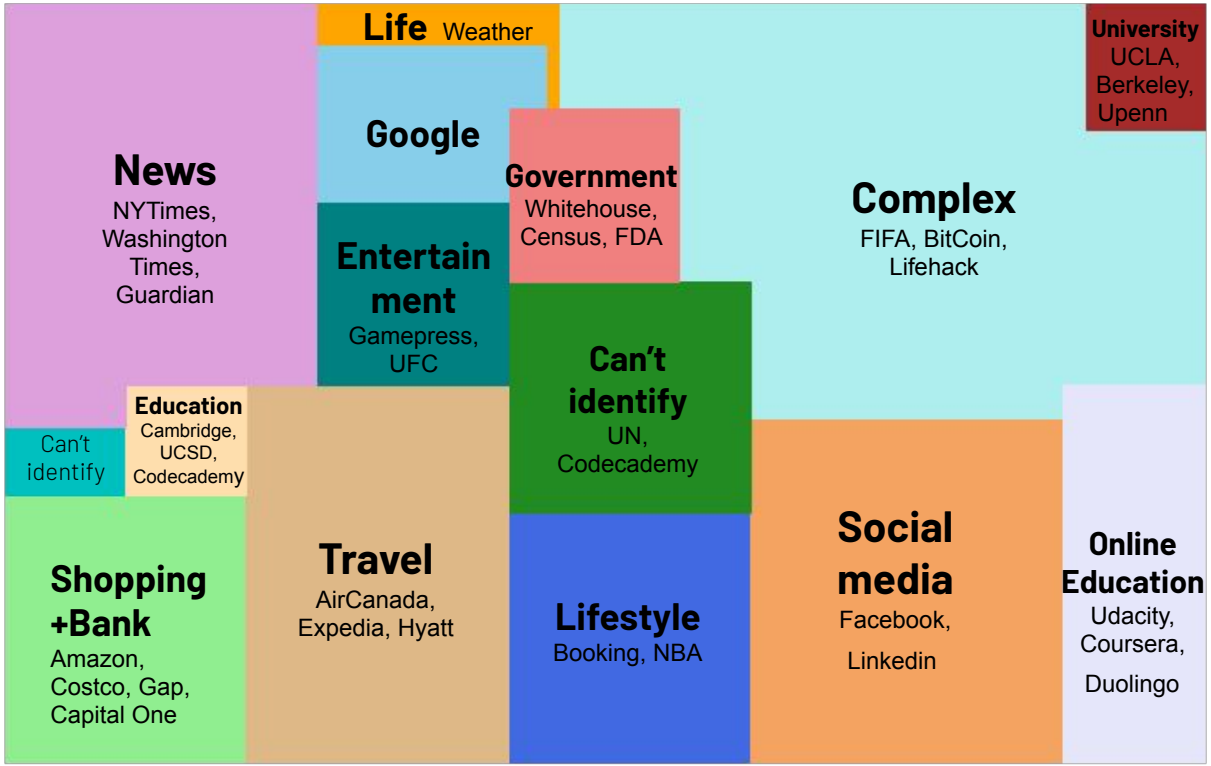
Modeling I

K-Means Clustering for Companies



Find **15** company clusters based on privacy policy corpus

Clustering Result



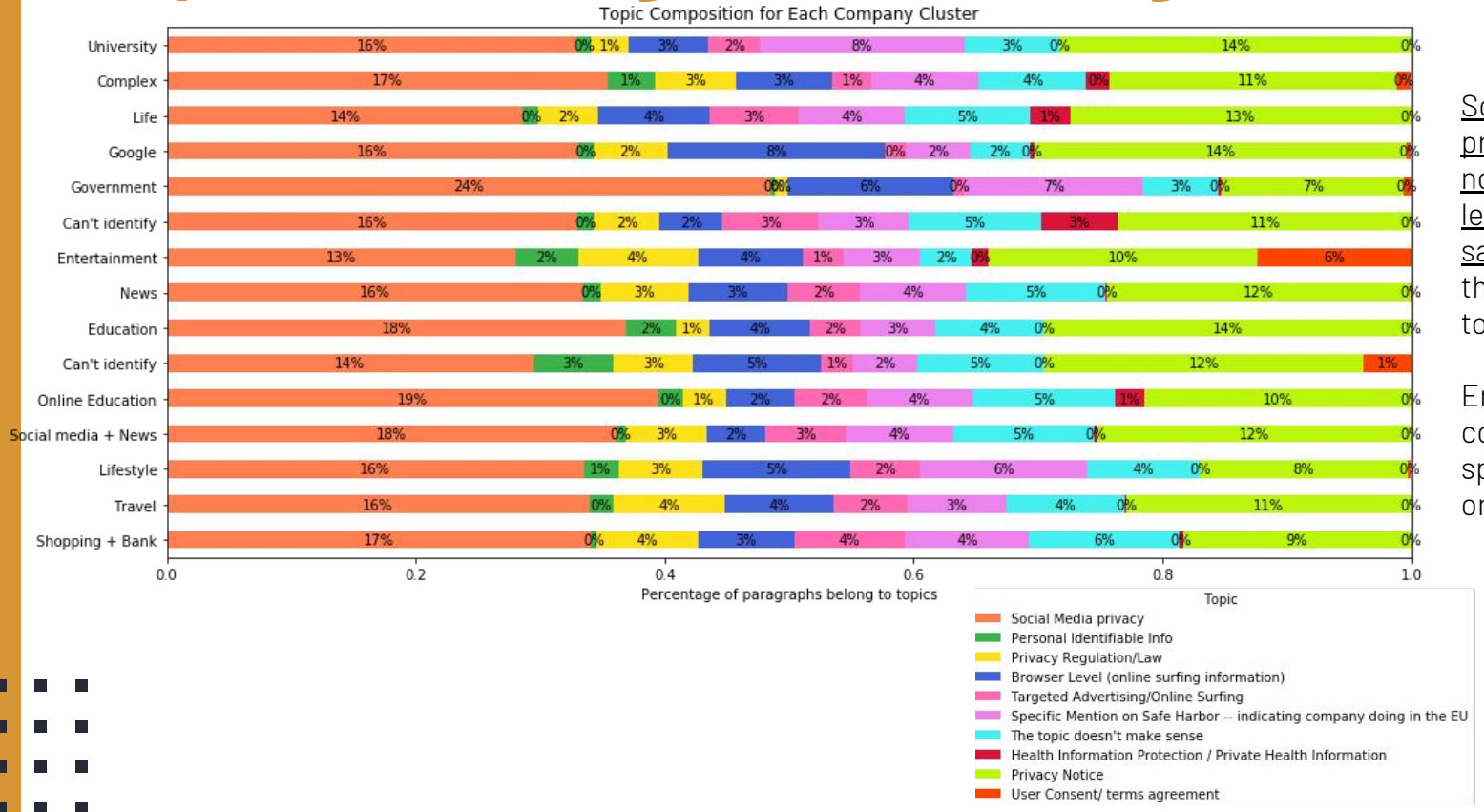
In most clusters, companies are in same industry or share similar business model.

Modeling II

Topic Modeling for Privacy Policies

- Applied LDA topic modeling with number of topics = 10 at paragraph level
- Asked information security professional to identify content under each topic

Topic Modeling & Clustering



Social media privacy, privacy notice, browser level information, safe harbor are the most popular topics.

Entertainment companies pay special attention on user consent.

Specific Topics

Social Media Privacy

information : 0.0985003888802235
use : 0.033832048355656064
collect : 0.03372265450108539
service : 0.02265998190616343
address : 0.02097776184439887
provide : 0.01811076233145754
site : 0.01728929539128226
personal : 0.015910479790198388
website : 0.013788870379000564
include : 0.013785837879824158
number : 0.0114830192358916
device : 0.011158839926674649
mobile : 0.010672397402423145
services : 0.009657927326876922
application : 0.009221924288049315
social : 0.009173015770771049
personally : 0.009005001360826161
datum : 0.008168949988182166
user : 0.007747083207592519
access : 0.0077199875323793465

Privacy Notice

policy : 0.08928033861104702
privacy : 0.08844273327439488
change : 0.04070734286376891
use : 0.040457373286114935
information : 0.036643714256386684
time : 0.0197954014994107
term : 0.016366011302038047
site : 0.015538053045746547
notice : 0.01462332058409758
website : 0.01428967732860761
united : 0.014086749600000273
update : 0.013604056485145627
states : 0.012837938499449543
post : 0.012377584548007473
personal : 0.010991480733106913
consent : 0.010921043456799277
collect : 0.010526584557379228
service : 0.008834925505062648
provide : 0.008486221591227831
transfer : 0.008064796257548301

Specific Topics

Browser Information

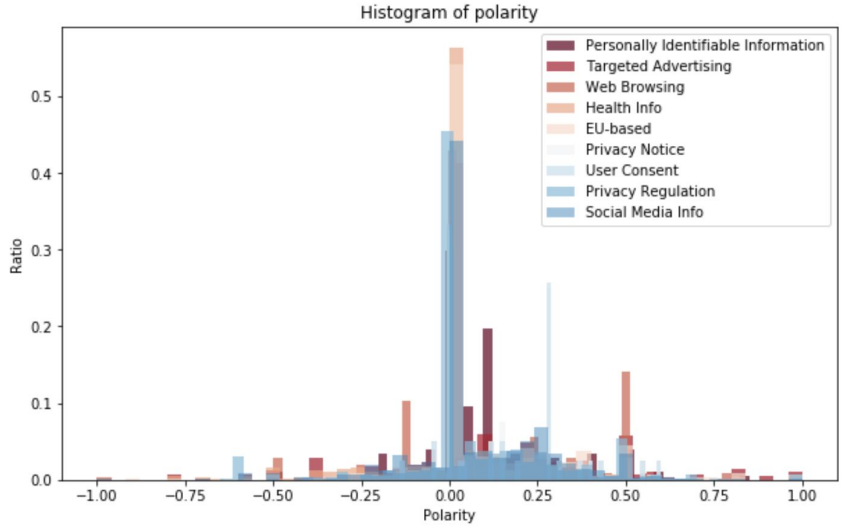
cookie : 0.06099875550840644
use : 0.043310365032977007
site : 0.02792040921425496
information : 0.026958475875924766
web : 0.0249343144178335
browser : 0.02035237026596797
website : 0.01819643668222169
party : 0.01736430883017555
advertising : 0.017329169990731745
visit : 0.015319384826398703
ad : 0.015103020440062582
user : 0.010591703093760016
computer : 0.010517337367098253
service : 0.010499954698462198
advertisement : 0.009949031088480957
page : 0.009713778251209759
opt : 0.00880011518585138
collect : 0.00828248421533964
company : 0.007652432461079144
technology : 0.007531461531618216

Safe Harbor

privacy : 0.08052999496241008
information : 0.04386154126418942
policy : 0.0438296533330121
site : 0.03354028232473401
website : 0.028462168252934458
com : 0.025129970288863032
personal : 0.01907002980717302
link : 0.018268743564421153
protect : 0.015212596495309856
collect : 0.015026813716442617
use : 0.01499750149949308
party : 0.01424188109489105
contact : 0.014002229505402437
web : 0.011604293835032586
statement : 0.010868812065618284
question : 0.009711910965488575
safe : 0.009386878904212984
www : 0.008542965385821144
truste : 0.008271226225508087
harbor : 0.007906276248752998

Sentiment Analysis on Topics

- **Sentiment** [Polarity Score]



Negative **Neutral** **Positive**

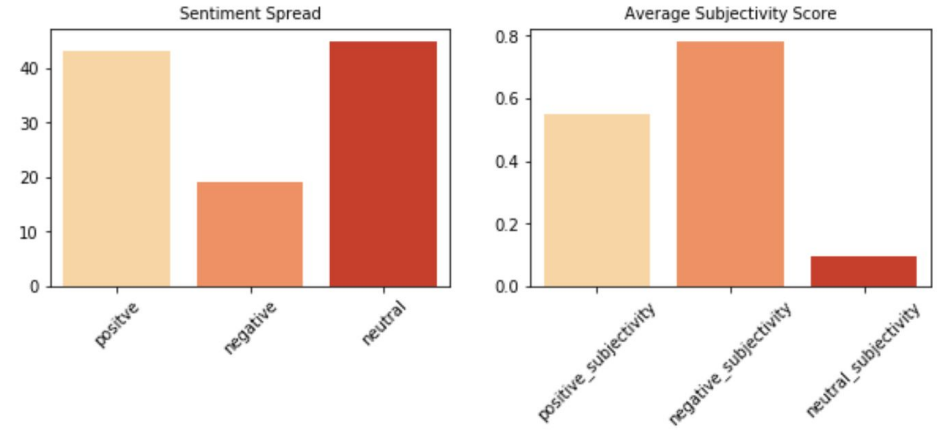
Most tweets are neutral across all topics

There are more positive sentiment than negative across all topics

Sentiment Analysis Cont.

- **Subjectivity** [Public Opinion - Factual Information]

Web Browsing

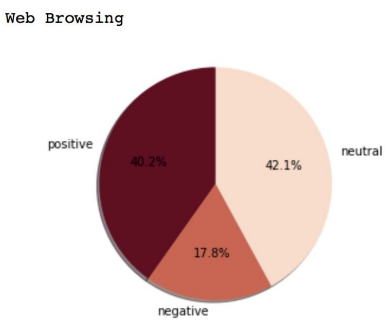


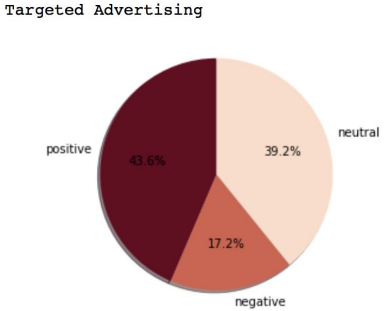
Subjectivity tells the scale of the sentiment

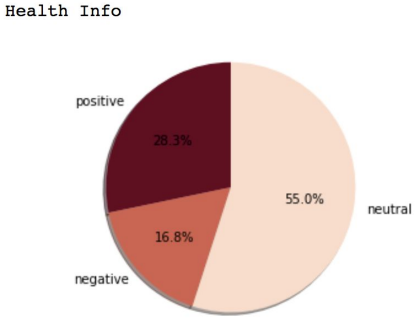
For example, the scale of negative sentiment towards Web Browsing Information is stronger than positive sentiment

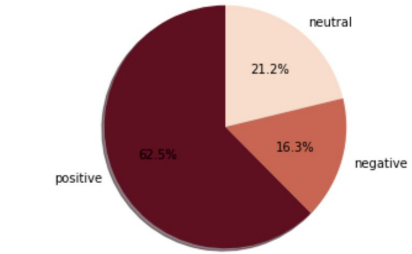
Sentiment Analysis Cont.

- Top 4 Privacy Topics with High Negative Sentiment









Web Browsing Information

Targeted Advertising

Social Media

Health Information

Sentiment - Cluster Mapping

Health Information

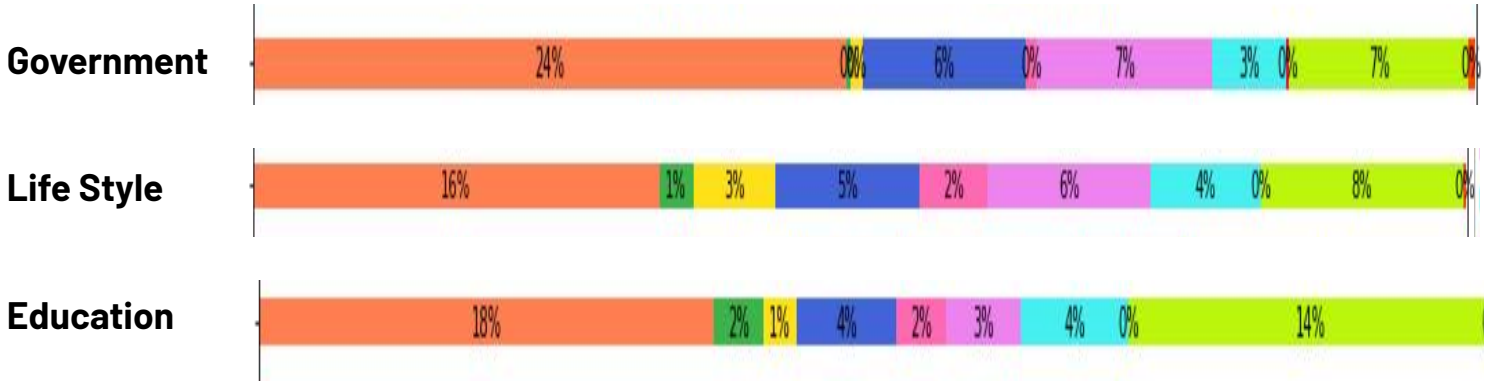
Social Media

Browser level Info

Personally Identifiable Information

4 most concerned topics

3 mapped company cluster



Policy Recommendation

Policy Recommendation

Our results correspond to the Pew Research survey , in which **59%** people feel that there is a **lack of understanding** of data that is collected from Companies and **78%** from government. Therefore, we suggest the government to implement below policies and regulations:

Short-term

- The government, as a regulator, should confront the distrust and concern from the public and **highlight its data privacy documents** accordingly.
- **Strengthen supervision** on companies whose data privacy policy containing most concerned topics.

Long-term

- Categorizing companies by data privacy focus and offer **baseline regulations** for each category.
- Enforce a transparent and accessible **information collecting process** for everyone.
- Empower the general public by Increasing **education** on data privacy law.

Thank you
Q&A