Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data

Hao Zhang[1], Gunhee Kim[2], Eric P. Xing[1]

[1]: Carnegie Mellon University, [2]: Seoul National University

Problem Statement

The increasing pervasiveness of Internet has lead to a wealth of consumer-created data over a multitude of online platforms

What can we learn?

- General public's opinion towards different companies' products and service
- Performance evaluations in different market conditions (time, location etc.)

Brand Competitions

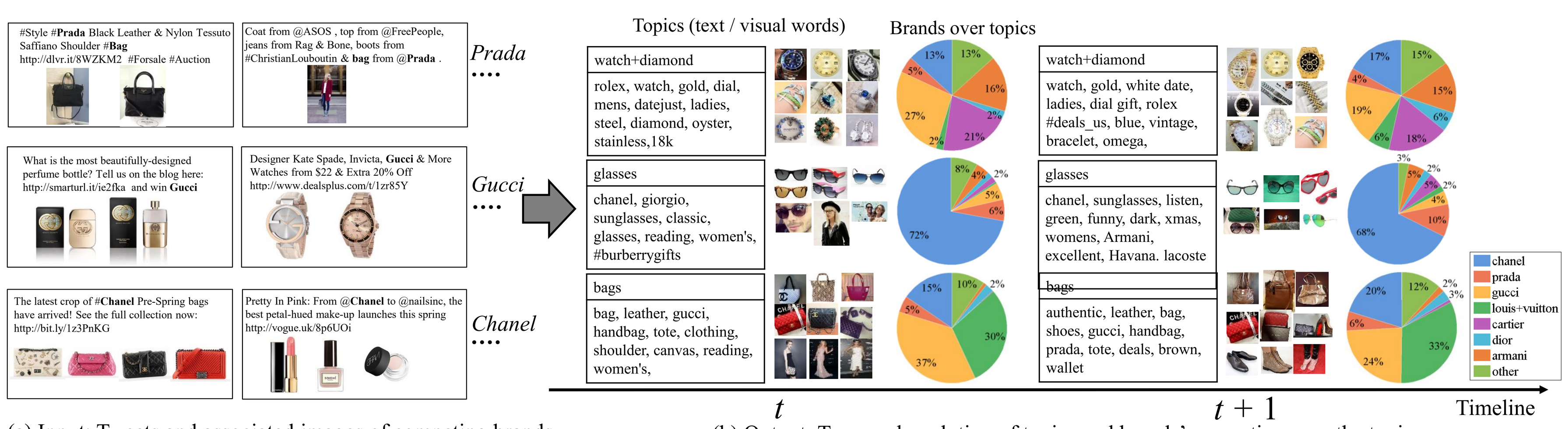
What does marketers want to see?

- **Detection:** Listen in consumers' opinions towards their products and their competitors
- **Summarization:** Summarize/visualize how a shared market is occupied by different brands
- **Dynamics:** Monitoring the changes of market competition over time

Our Approach: Joint Analysis of Text and Images

Take advantage of the pervasiveness of images on the social media

- A large portion of tweets simply show images&links without any meaningful text in them. Images play an important role for representing topics in this type of documents
- Many users prefer to use images to deliver their idea more clearly and broadly,
- The joint use of images with text also helps marketers interpret the discovered topics
- Images may be essential for users to make conversation about customers' descriptions, experiences, and opinions toward the brands.

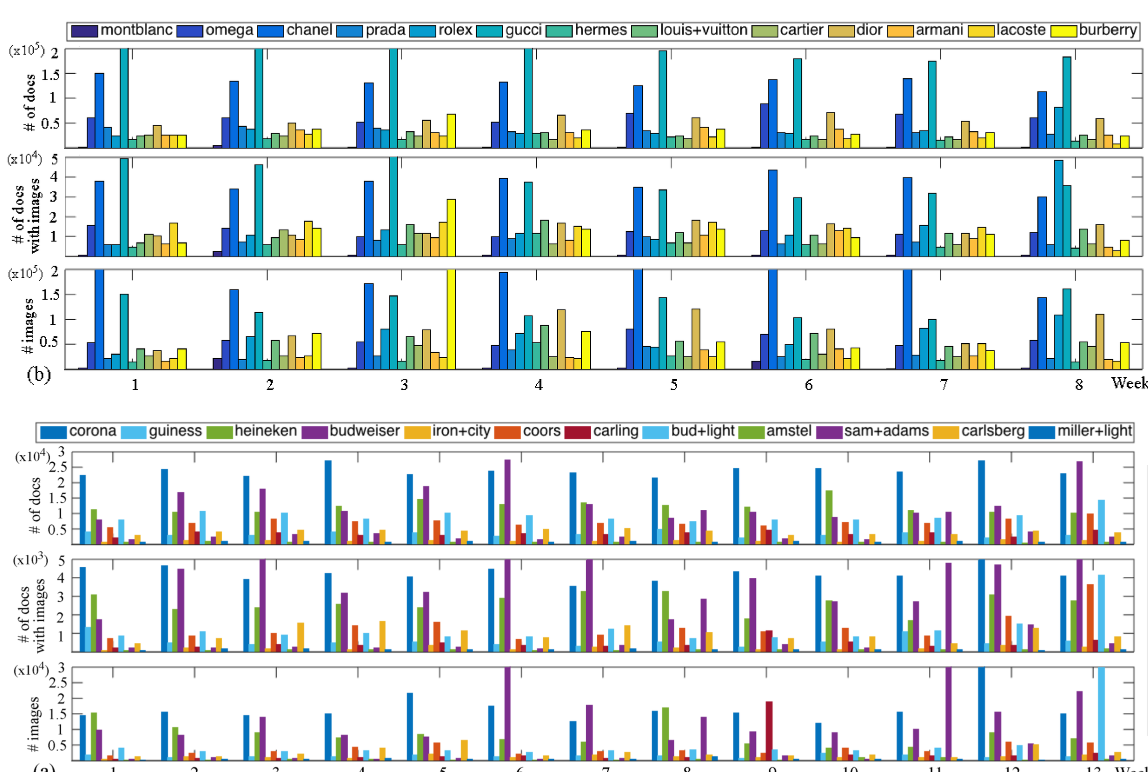


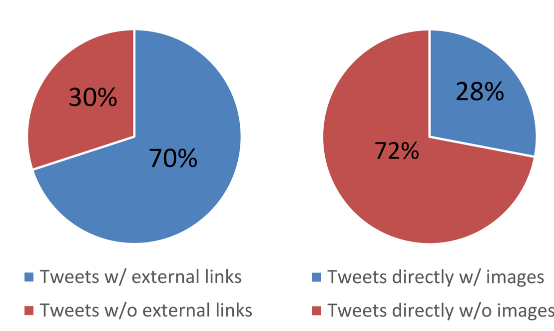
(a) Input: Tweets and associated images of competing brands (b) Output: Temporal evolution of topics and brands' proportion over the topics

Collecting Data

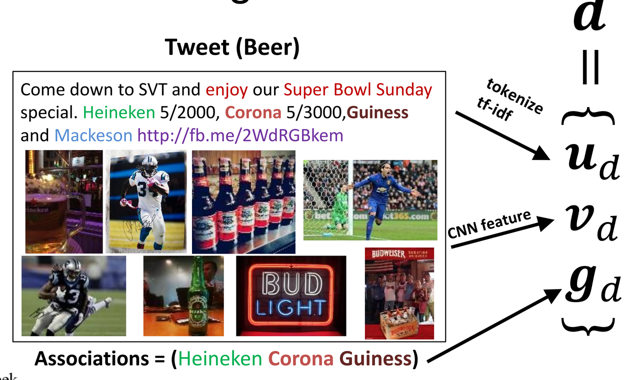
Crawling raw tweets and associated Images using the **REST** API

- 2 groups of brands: **Luxury** (13 brands) **Beer** (12 brands)
- **6.6M** tweets and **7.5M** images from twitter and external links
- Time range: **10/20/2014** to **02/01/2015**





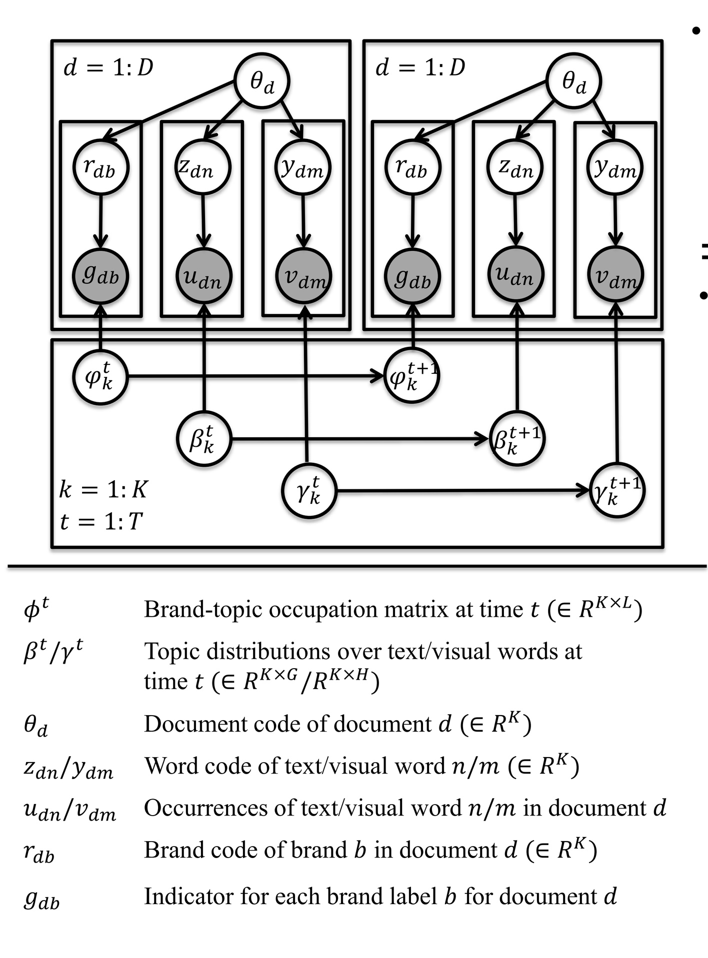
Processing



Competitive Dynamic Multi-view STC (cdSTC)

The model aims to address 3 major challenges

1. Multi-view — Modeling of multi-view representations of text and images
2. Competition — Modeling of latent topics that are competitively shared by multiple brands
3. Dynamic — Tracking temporal evolution of the topics and brand competitions



- Joint Probability

$$p(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{u},\boldsymbol{y},\boldsymbol{v},\boldsymbol{r},\boldsymbol{g}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi}) = p(\boldsymbol{\theta})\prod_{n\in N} p(z_n|\boldsymbol{\theta})p(u_n|z_n,\boldsymbol{\beta})\prod_{m\in M} p(y_m|\boldsymbol{\gamma})p(v_m|y_m,\boldsymbol{\gamma}) \prod_{b\in B} p(r_b|\boldsymbol{\phi})p(g_b|r_b,\boldsymbol{\phi})$$

- MAP

$$\min_{\{\boldsymbol{\theta}^t,\boldsymbol{\beta}^t,\boldsymbol{\gamma}^t,\boldsymbol{\phi}^t\}_{t=1}^T} \sum_{t=1}^{T}\sum_{d=1}^{D} \lambda\|\boldsymbol{\theta}_d^t\|_1 + \sum_{t=1}^{T}(\pi_1\|\boldsymbol{\beta}^t-\boldsymbol{\beta}^{t-1}\|_2^2 + \pi_2\|\boldsymbol{\gamma}^t-\boldsymbol{\gamma}^{t-1}\|_2^2 + \pi_3\|\boldsymbol{\phi}^t-\boldsymbol{\phi}^{t-1}\|_2^2)$$
$$+ \sum_{t=1}^{T}\sum_{d=1}^{D^t}\sum_{n\in N_d^t}(\nu_1\|\boldsymbol{z}_{dn}^t-\boldsymbol{\theta}_d^t\|_2^2 + \rho_1\|\boldsymbol{z}_{dn}^t\|_1 + L(\boldsymbol{z}_{dn}^t,\boldsymbol{\beta}^t))$$
$$+ \sum_{t=1}^{T}\sum_{d=1}^{D^t}\sum_{m\in N_d^t}(\nu_2\|\boldsymbol{y}_{dm}^t-\boldsymbol{\theta}_d^t\|_2^2 + \rho_2\|\boldsymbol{y}_{dm}^t\|_1 + L(\boldsymbol{y}_{dm}^t,\boldsymbol{\gamma}^t))$$
$$+ \sum_{t=1}^{T}\sum_{d=1}^{D^t}\sum_{b\in B_d^t}(\nu_3\|\boldsymbol{r}_{db}^t-\boldsymbol{\theta}_d^t\|_2^2 + \rho_3\|\boldsymbol{r}_{db}^t\|_1 + L(\boldsymbol{r}_{db}^t,\boldsymbol{\phi}^t))$$
$$s.t.\ \boldsymbol{\theta}_d^t > 0, \forall d, t. \boldsymbol{z}_{dn}^t, \boldsymbol{y}_{dm}^t, \boldsymbol{r}_{db}^t > 0, \forall d,n,m,b,t \quad \beta_k^t \in P_U, \gamma_k^t \in P_V, \phi_k^t \in P_B, \forall k,t$$

sparse term for document code; evolving chain; text; image; brand; simplex

ϕ^t	Brand-topic occupation matrix at time t ($\in R^{K\times L}$)
β^t/γ^t	Topic distributions over text/visual words at time t ($\in R^{K\times G}/R^{K\times H}$)
θ_d	Document code of document d ($\in R^K$)
z_{dn}/y_{dm}	Word code of text/visual word n/m ($\in R^K$)
u_{dn}/v_{dm}	Occurrences of text/visual word n/m in document d
r_{db}	Brand code of brand b in document d ($\in R^K$)
g_{db}	Indicator for each brand label b for document d

Evaluation: Topic Quality

amazon mechanical turk — Artificial Artificial Intelligence

Argument 1: Lower perplexity ≠ higher quality [J. Chang 2009]

Argument 2: Perplexity is not a fair metric for models with different distributions

–Define the **Coherence Measure (CM)** and the **Validity Measure (VM)**:

$$CM = \frac{\#\ of\ relevant\ words}{\#\ of\ words\ in\ valid\ topics} \qquad VM = \frac{\#\ of\ valid\ topics}{\#\ of\ topics}$$

•**Average VM/CM on text topics**

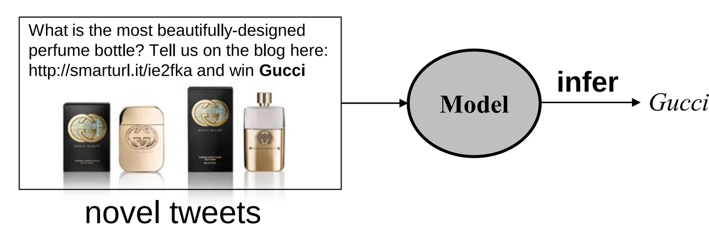
	VM (Beer / Luxury)	CM (Beer / Luxury)
dLDA	0.53 / 0.68	0.55 / 0.52
STC + dyn	0.44 / 0.66	0.57 / 0.57
cdSTC + multi	0.51 / 0.70	**0.63** / **0.59**
cdSTC + text	**0.605** / **0.71**	0.61 / **0.59**

•**Average VM/CM on visual topic**

	VM (Beer / Luxury)	CM (Beer / Luxury)
Kmeans	0.39 / 0.56	0.59 / 0.64
LDA + multi	**0.57** / 0.63	0.51 / 0.69
cdSTC + multi	**0.57** / **0.65**	**0.66** / **0.71**

Evaluation: Prediction

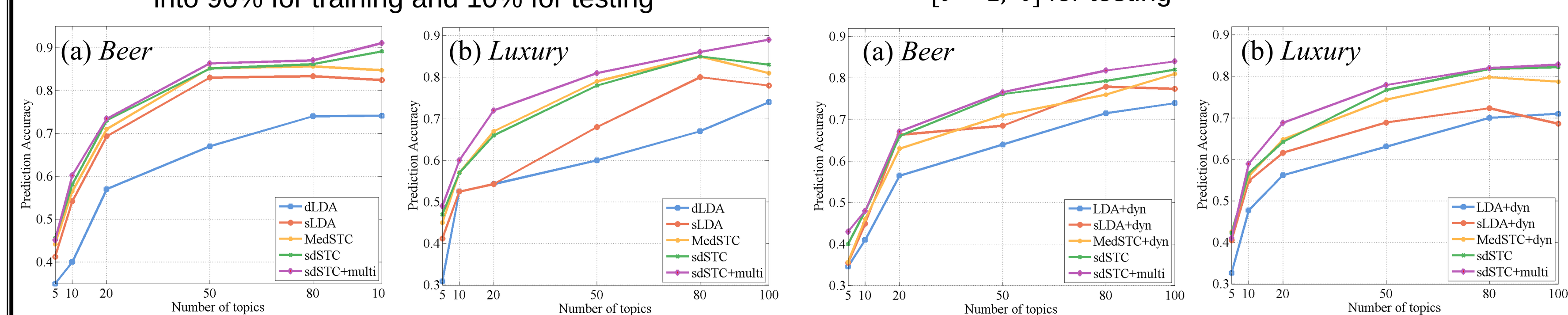
- **Task I**: Given a novel tweet, can we predict its most associated brand?



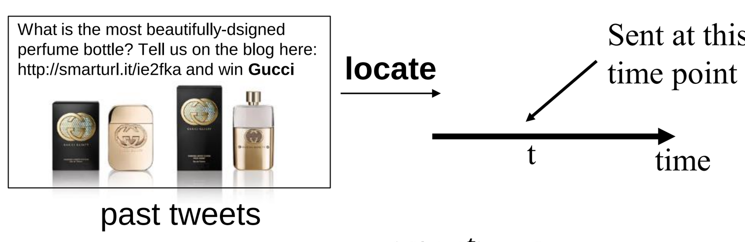
$$\max_{\{\boldsymbol{\theta}^t,\boldsymbol{\mathcal{M}}^t,\boldsymbol{\eta}^t\}_{t=1}^T} \sum_{t=1}^{T} f(\Theta^t,\boldsymbol{\mathcal{M}}^t,D^t) + CR(\Theta^t,\boldsymbol{\eta}^t) + \frac{1}{2}\|\boldsymbol{\eta}^t\|_2^2$$
$$s.t.\ \boldsymbol{\theta}_d^t > 0, \forall d, t. \boldsymbol{z}_{dn}^t, \boldsymbol{y}_{dm}^t > 0, \forall d,n,m,t \quad \beta_k^t \in P_U, \gamma_k^t \in P_V, \forall k,t$$

Task I-I: Randomly split data in every time slice into 90% for training and 10% for testing

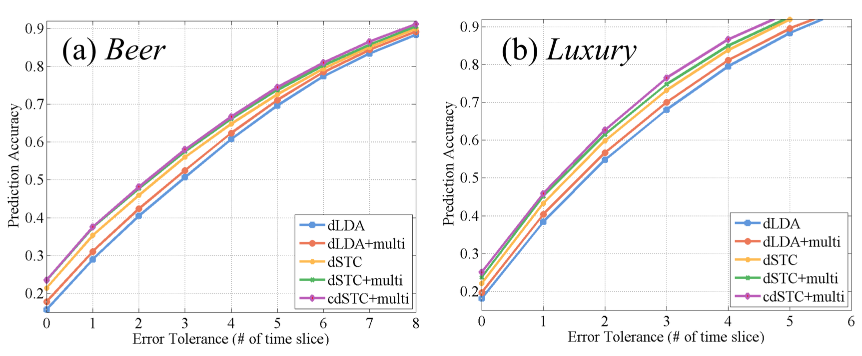
Task I-II: Use the data in $[1, t-1]$ for training, $[t-1, t]$ for testing



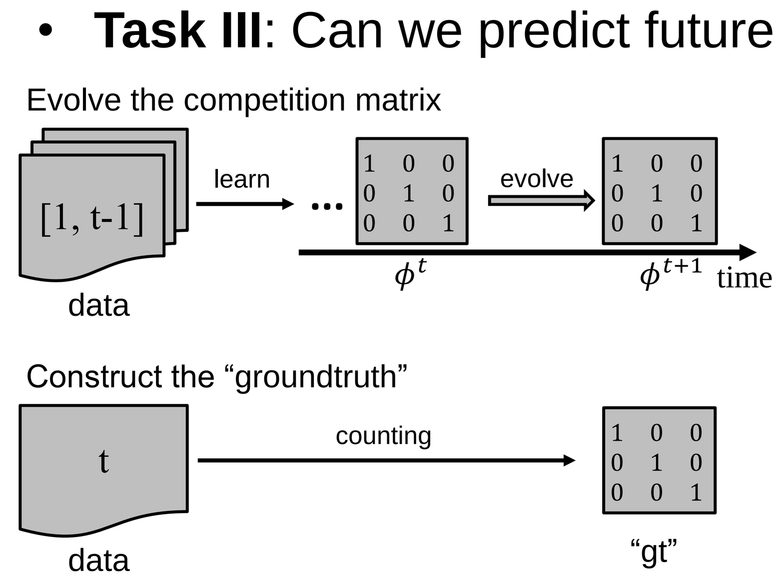
- **Task II**: Given an unseen past document, can we predict its timestamp?

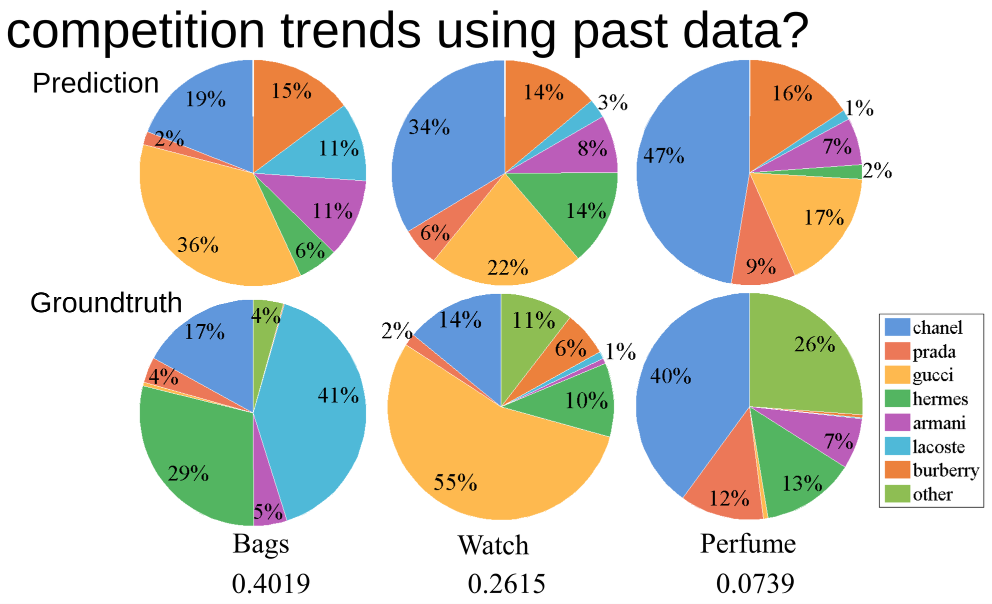


$$\max_t\ p(d|\boldsymbol{\mathcal{M}}^t), where$$
$$p(d|\boldsymbol{\mathcal{M}}^t) = \prod_{n\in N_d} p(u_n|\boldsymbol{\beta}^t)\prod_{m\in M_d} p(v_m|\boldsymbol{\gamma}^t)\prod_{b\in B_d} p(g_b|\boldsymbol{\phi}^t)$$



- **Task III**: Can we predict future competition trends using past data?





Brand Competition Monitoring

Objective

- How brands occupy the market in every time slice?
- How each textual/visual topic evolves over time?
- How each brand's occupation changes over time?
- How's the competition trends between multi-brands like over time?

easy → difficult