# Item2Vec

# Agenda

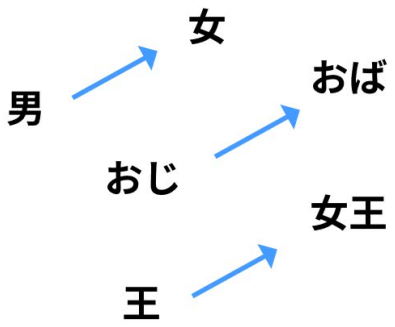
- Item2Vec(Word2Vec)
- Application, Dataset
- Experiments
- Results
- Wrap Up

# Item2Vec(Word2Vec)

**各Wordを0,1ではなく、ベクトル化することで、意味をもたせる表現方法**


特徴が捕らえられるので、意味に基づいた関係性が見えてくる

- 「王様」- 「男」＋ 「女」＝ 「女王」
- 「パリ」- 「フランス」＋ 「日本」＝ 「東京」




**Item2Vec**

このWord2Vecの方法をItemに応用し、Recommendation等に応用しているもの

# Application, Dataset

今回、「テレビドラマ」に注目してみた

→ 以下のPaperよりデータセット引用

## 放送前の情報のみを用いたテレビドラマの視聴率予測

Predicting Ratings of Japanese TV Dramas Based on Cast and Popularity

学生会員　福 島 悠 介†，　正会員　山 崎 俊 彦†，　正会員　相 澤 清 晴†

Yusuke Fukushima†, Toshihiko Yamasaki† and Kiyoharu Aizawa†

表 2　視聴率に影響を与えた上位 25 要因

| 特徴量 | 重み |
|---|---|
| 放送回数 25 回以上 | +2.90 |
| 木村拓哉（俳優） | +2.70 |
| DREAMS COME TRUE（主題歌） | +2.12 |
| 阿部寛（俳優） | +2.08 |
| 午後 8 時開始 | +2.02 |
| 米倉涼子（女優） | +2.00 |
| フジテレビ，月曜，午後 9 時開始 | +1.98 |
| NHK，日曜，午後 8 時開始 | +1.88 |
| 水谷豊（俳優） | +1.85 |
| 午後 10 時終了 | +1.77 |
| 森本梢子（原作者） | +1.64 |
| Superfly（主題歌） | +1.63 |

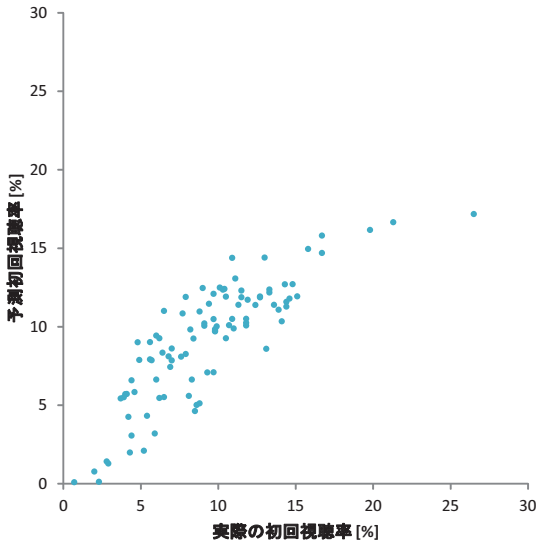



図 6　全特徴量を用いた際の実際の視聴率と予測視聴率

2008年4月から2015年6月の間のドラマ

678本

特徴量

- キャスト
- 放送時間
- 放送回数
- 視聴率
- 主題歌歌手
- 演出家
- 作曲家
- 脚本家

# Application, Dataset

■期待していたこと、こんなことが見れたりするか！！

木村拓哉 – 高視聴率 + 低視聴率 = ○○？

新垣結衣– ゴールデンタイム + 深夜 = ○○？

二宮和也 – 嵐 + SMAP = 草なぎ剛？

# Experiments

## ■データの準備

・データを、各ドラマの「平均視聴率、メインキャスト２名、主題歌歌手、放送時間帯」の５つに絞り、それを１セットとする
・視聴率は、５％区切りでグループ化する
・キャスト、歌手は、１回しか登場しないものはOtherCast, OtherSingerとしてまとめる
・上記作業で、対象項目は388

ココがいまいちポイント

登場回数

放送時間帯

"18": 10,
"19": 36,
"20": 68,
"21": 236,
"22": 175,
"23": 87,
"24": 51,
"25": 11,
"26": 4,

平均視聴率

"under5": 123,
"5~10": 282,
"10~15": 219,
"15~20": 43,
"20~25": 10,
"over25": 1,

キャスト(10回以上)

"渡瀬恒彦": 12,
"仲間由紀恵": 12,
"上戸彩": 11,
"佐々木蔵之介": 10,
"沢村一樹": 10,
"観月ありさ": 10,

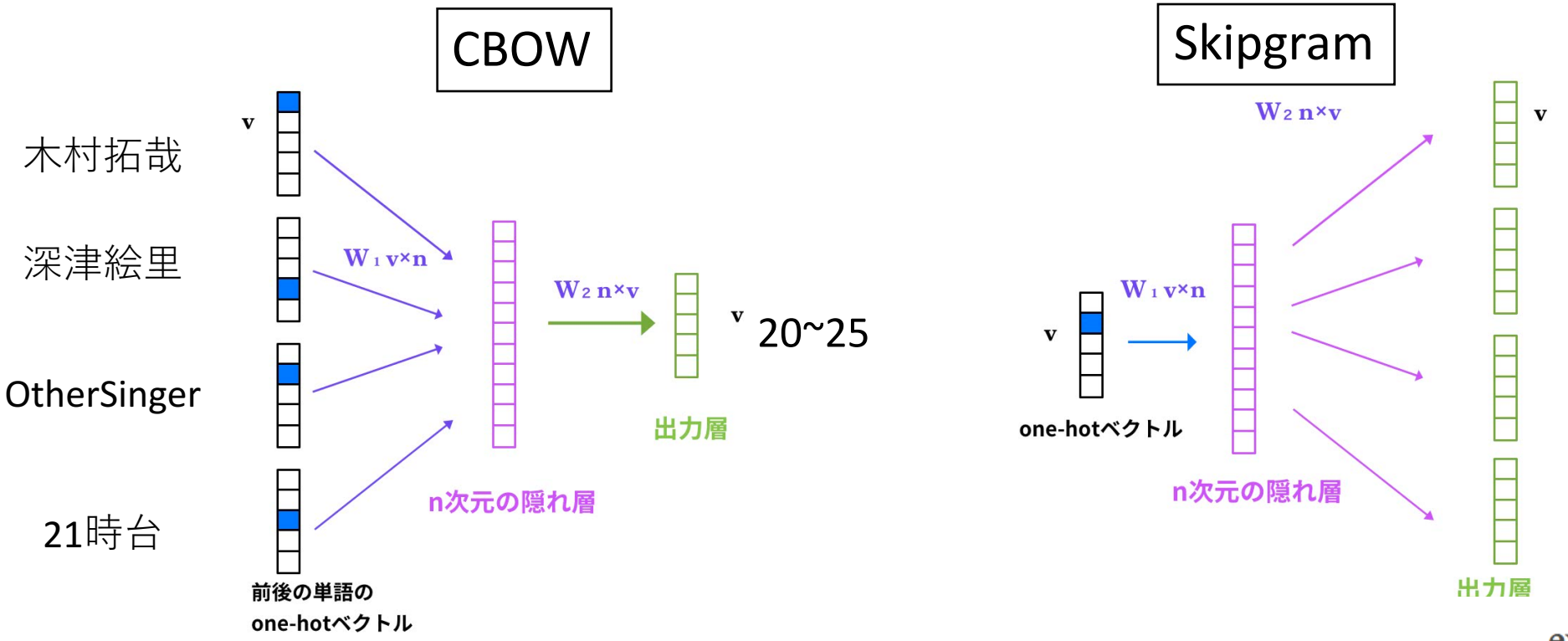
歌手(10回以上)

"None": 171,
"嵐": 17,
"関ジャニ": 10,

"otherCast": 216,
"otherSinger": 178,

# Experiments

■ネットワーク、実験



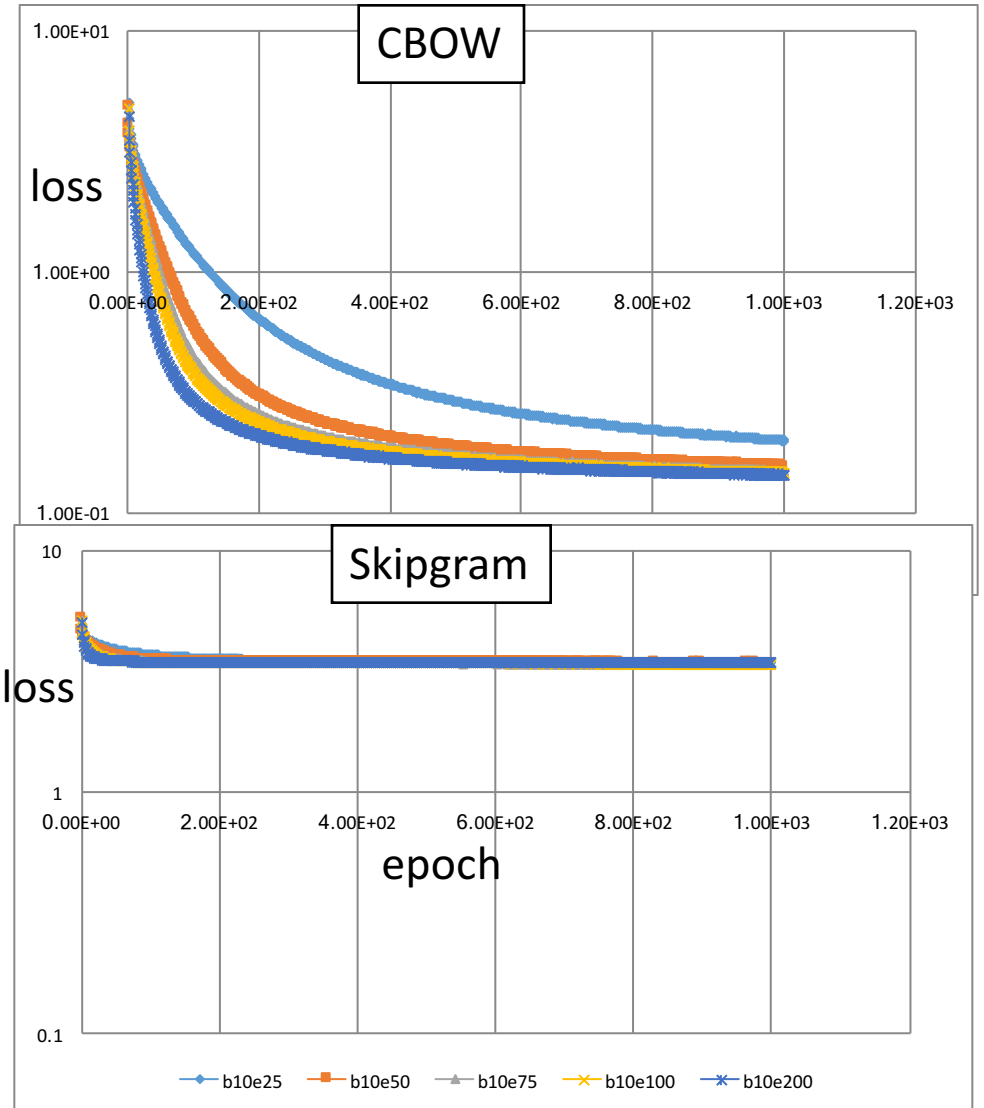

$$p(w_O|w_I) = \frac{\exp(\boldsymbol{v}'^{\mathrm{T}}_{w_O} \cdot \boldsymbol{v}_{w_I})}{\sum_{w_v \in V} \exp(\boldsymbol{v}'^{\mathrm{T}}_{w_v} \cdot \boldsymbol{v}_{w_I})}$$

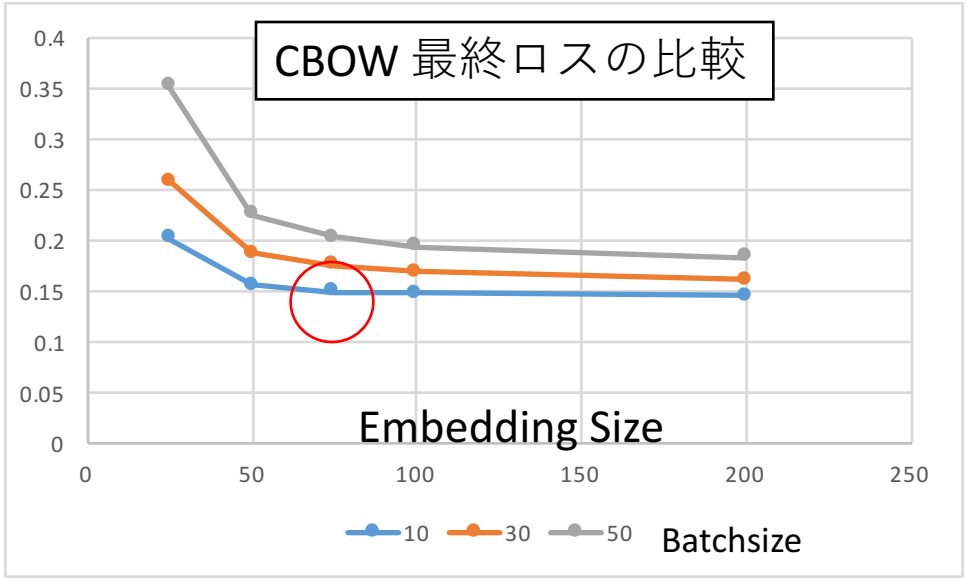
Softmax

- 各ドラマの5つの項目の穴埋め問題として計算
- CBOW, Skipgram両方試した
- Batchsize [10,30,50] , Embeddingサイズ [25, 50, 75, 100, 200]
- 1000 epoch
- ネガティブサンプリングなし

# Results

## ■Training



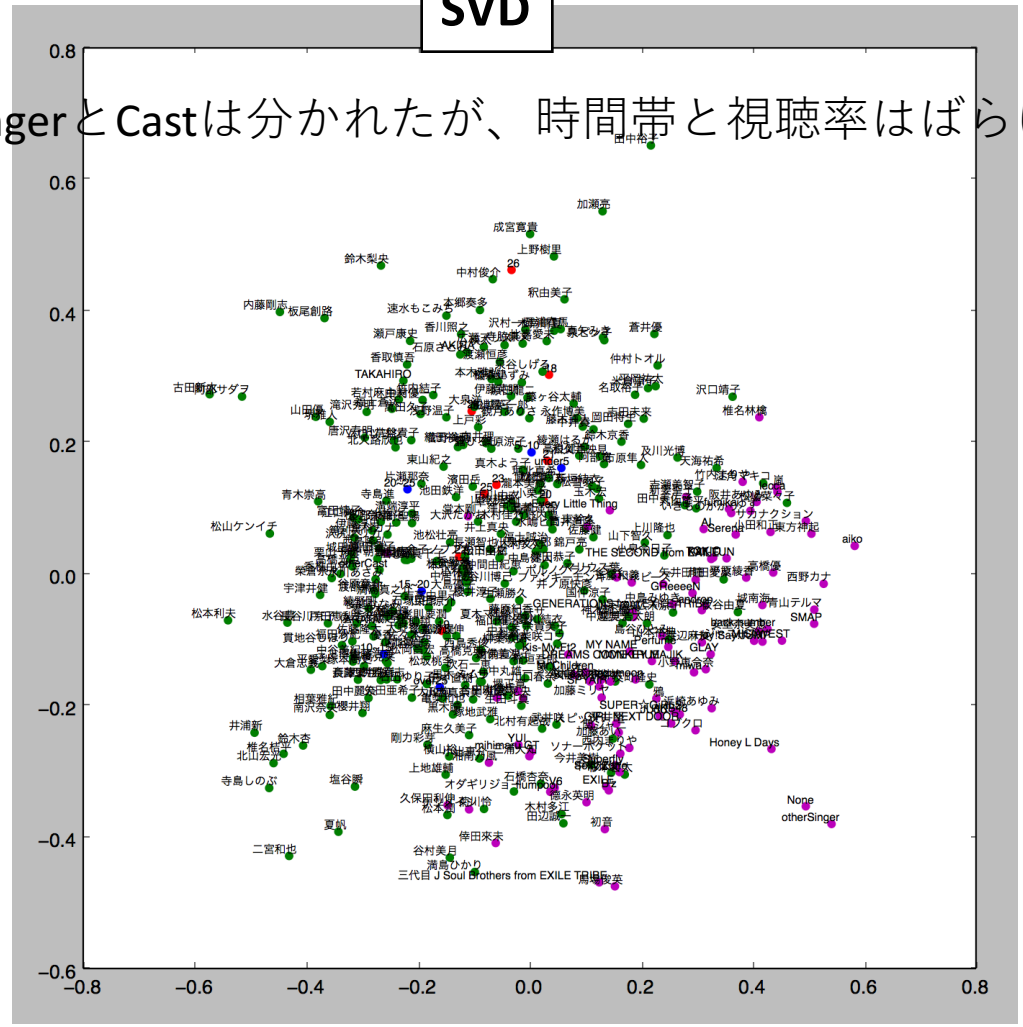




- CBOWの方が明らかにロス小さい
- Batchsize小さい方が収束早い
- Embedsize大きい方が収束早い
- CBOW, Batchsize:10, Embedsize:75を最適と判断

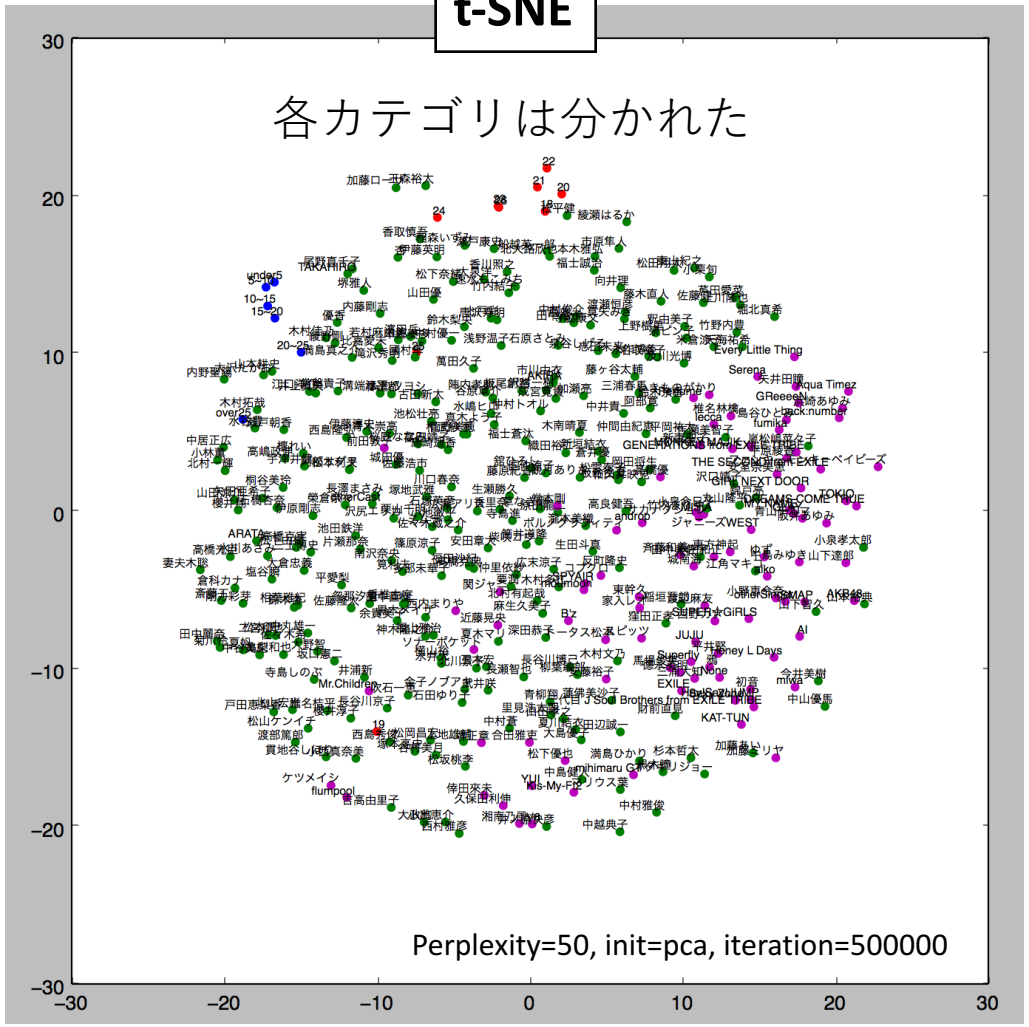
# Results

## ■2d-visualize

**SVD**



**t-SNE**




SingerとCastは分かれたが、時間帯と視聴率はばらけている

各カテゴリは分かれた

t-SNEはHyperparameterの影響を受けやすい、また繰り返すだけでも変化が大きく、厄介！

# Results

## ■Similarity

75次元のEmbeddingベクトルから、コサイン距離で最も近いものをリスト化

| Nearest to "二宮和也" | Nearest to "錦戸亮" | Nearest to "木村拓哉" | Nearest to "草なぎ剛" | Nearest to "新垣結衣" | Nearest to "上戸彩" |
|---|---|---|---|---|---|
| "夏帆" | "丸山隆平" | "水谷豊" | "木村多江" | "蒼井優" | "速水もこみち" |
| "松本潤" | "androp" | "北大路欣也" | "若村麻由美" | "成宮寛貴" | "蒼井優" |
| "櫻井翔" | "田中美佐子" | "over25" | "堂本剛" | "長澤まさみ" | "佐藤浩市" |
| "相葉雅紀" | "蒼井優" | "釈由美子" | "大泉洋" | "仲間由紀恵" | "前田敦子" |
| "鈴木杏" | "木南晴夏" | "本木雅弘" | "井上真央" | "嵐" | "稲森いずみ" |
| "水谷豊" | "MISIA" | "藤ヶ谷太輔" | "山田涼介" | "吉高由里子" | "石原さとみ" |
| "北村一輝" | "小田和正" | "松田翔太" | "矢田亜希子" | "錦戸亮" | "沢尻エリカ" |

グループ感が出ている　木村拓哉と相棒　女優陣多し　女優でも種類が違う？

Similarityはそんなに的外れではなさそう

# Results

## ■Arithmetic

75次元のEmbeddingベクトルで足し引きしたベクトルの、コサイン距離で最も近いものをリスト化

| | | | |
|---|---|---|---|
| "木村拓哉" | "新垣結衣" | "二宮和也" | "二宮和也" |
| minus | minus | minus | minus |
| "over25" | 22 | "嵐" | "嵐" |
| plus | plus | plus | plus |
| "under5" | 25 | "SMAP" | "関ジャニ" |
| --- | --- | --- | --- |
| "under5" | "新垣結衣" | "二宮和也" | "二宮和也" |
| "木村拓哉" | 25 | "SMAP" | "関ジャニ" |
| "余貴美子" | "吉高由里子" | "櫻井翔" | "相葉雅紀" |
| "及川光博" | "竹内結子" | "松本潤" | "櫻井翔" |
| "釈由美子" | "仲里依紗" | "田中麗奈" | "長谷川京子" |
| "上川隆也" | "上戸彩" | "北村一輝" | "松本潤" |
| "真矢みき" | "仲間由紀恵" | "AKB48" | "北山宏光" |

75次元だと、上位2位がなぜかBaseとPlus要素
想定していたような結果は出ず

# Results

## ■Validation

実際の学習済みのCBOWの結果を用いて、夢の共演をさせてみる

```
input
[
    "堺雅人",
    "上戸彩",
    21,
    "None"
]
prediction
"over25"
0.988974
"10~15"
0.00940624
"木村拓哉"
0.00151455
```

```
input
[
    "木村拓哉",
    "新垣結衣",
    21,
     "Mr.Children"
]
prediction
"15~20"
1.0
"20~25"
2.27839e-07
"山下智久"
3.49408e-08
```

```
input
[
    "木村拓哉",
    "草なぎ剛",
    21,
     "SMAP"
]
prediction
"10~15"
0.999981
"20~25"
1.28996e-05
"今井美樹"
2.6446e-06
```

半沢直樹の計算は正しい

# Wrap Up

- Item2Vec(Word2Vec)の技術をテレビドラマに応用してみた
- キャスト、主題歌歌手、時間帯、視聴率を基に解析して見て、各評価を実施して見た。想定ほどの面白い結果は残念ながら得られず。

- ネットワーク自体は非常にシンプルで実装は簡単（データ処理の方が大変）
- t-SNEによる2次元表現は、パラメータの影響が多すぎて、どれが正解かを見つけるのが結構難しい。
- Arithmeticも、元のベクトルのままではうまくできなかった