# Topic Clustering

## - Enron dataset (Lay, Skilling, Forney, Delainey)

## - Training:Test in 9:1

**Fech Scen Khoo**
Universität des Saarlandes

23 November 2020

## Model

"Measuring the diffusion of innovations with paragraph vector topic models" (Lenz & Winker, PLOS ONE (2020)):

paragraph vector + Gaussian mixture model

▷ paragraph vector: document representation
↪ PV-DBOW: Given the paragraph vector, predict the words in the text window ($\sim$ Skip-Gram model)

▷ Gaussian mixture model: soft-clustering version of k-means $\rightsquigarrow$ A sample can be assigned to more than one clusters
↪ 4 types of covariances in the Gaussian density components: diagonal, spherical, full,
tied (all components share the same covariance)

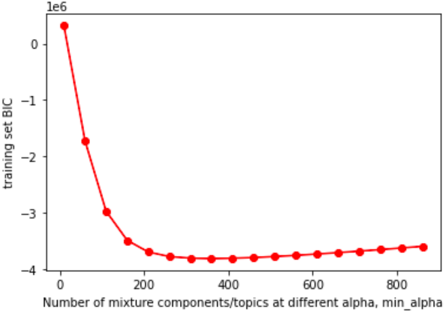
## Model parameters

Parameters for Doc2Vec:
*train word vectors simultaneously with doc vector*
dimensionality of the feature vectors, $d$
epochs, $e$
initial learning rate, $\alpha$
final learning rate, $min\,\alpha$
negative sampling, *neg*
downsampling of higher frequency words, *sample*

Parameters for GMM:
number of Gaussian components, $n$
covariance type, *cov*
*covariance regularization*

## Optimal parameter setting

After text preprocessing:
$n \in [10, 900], cov = tied$
$d = 85, e = 6,\ neg = 5, sample = 10^{-3},\ \alpha = \frac{1}{n},$
$min\,\alpha = \frac{0.0001n}{\ell}$ where number of texts, $\ell = 5805$



$\Rightarrow$ 18 models

## Inference using the test documents

Infer a vector using the trained models (Doc2Vec part)

$+$

Predict posterior probability (GMM part)

$\Rightarrow$ Under each trained model, for each test document, get its predicted $n$ components with the highest probability

$\Rightarrow$ Gather the test documents in terms of $n$ predicted
e.g. For a trained model of $n = 10$,
group *i*: $\{$test docs $\mathcal{A}$ of predicted $n = 1, 2\}$,
group *ii*: $\{$test docs $\mathcal{B}$ of predicted $n = 3, 4\}$,
......

## Determination of *n* for word cloud realization

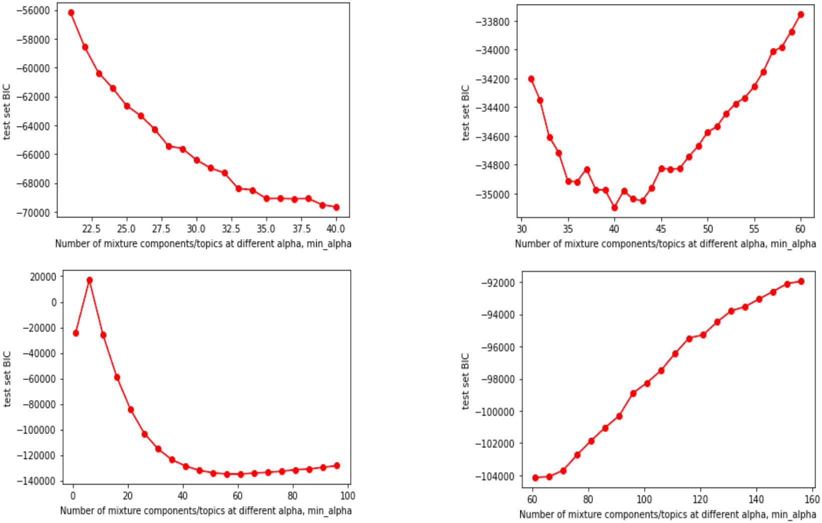
For each group of the test documents

provided that the number of docs $\geq$ largest $n$ in the range

plot the test set BIC wrt its range of $n$

make the word clouds at the $n$ with the lowest BIC value

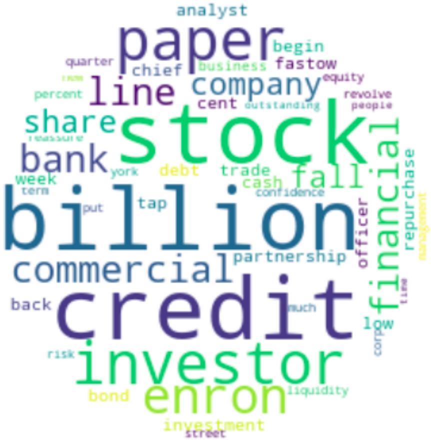
# Behaviours of the test BIC



*But, watch out for the range of n*

## A cloud from plot (1,1), Topic A

139 docs [trained model n160], BIC $\approx -68828$, $n = 36$



Top 15 central words: credit, billion, stock, investor, paper, enron, commercial, bank, line, financial, company, fall, share, partnership, begin

## Topics from plot (1,1)

139 docs [trained model n160], BIC $\approx -68828$, $n = 36$

Topic A: credit, billion, stock, investor, paper, enron, commercial, bank, line, financial, company, fall, share, partnership, begin

Topic B: incredible, reunion, everyone, future, forget, big, consider, tent, crowd, connect, pick, stay, year, far, ago

Topic C: asset, equity, percent, york, dollar, shareholder, decision, limit, ljm, partner, approval, security, news, lose, quarter

Topic D: width, spin, td, tr, table, class, br, border, cellspacing, cellpadding, normaltext, top, tag, information, right

## Topics from plot (2,1)

248 docs [trained model n260], BIC $\approx -134886$, $n = 60$

Topic A: tr, td, width, spin, table, class, border, cellpadding, cellspacing, normaltext, br, **gif**, member, information, top

Topic B: residential, llc, jimmie, show, loss, retail, hydrocarbon, gray, perform, propose, direction, high, entry, create

Topic C: story, oct, caracas, petroleumworld, nov, oil, venezuela, opec, previous, president, full, war, production, cut, ohep

Topic D: market, lehman, buy, company, expect, brother, old, inc, perform, trade, offer, estimate, business, strong, percentage

Topic E: natural, gas, business, news, power, electric, com, utility, co, trade, customer, new, chief, president, focus

Glossary: llc: limited liability company; opec: org. petroleum exporting countries; ohep: office of home energy programs

## Topics from plot (1,2)

75 docs [trained model n310], BIC $\approx -35056$, $n = 40$

Topic A: billion, credit, line, commercial, paper, investor, stock, bank, low, bond, financial, enron, company, tap, cent

Topic B: story, oct, petroleumworld, caracas, nov, oil, venezuela, opec, previous, president, full, elio, law, war, ohep

Topic C: expect, market, brother, lehman, point, percentage, buy, perform, estimate, inc, old, strong, street, eps, rate

Topic D: panel, jones, minute, dow, panelist, bandwidth, allocate, representative, follow, remark, session, energy, speaker, newswires, associate

Topic E: alliance, datum, sound, utility, online, energy, industry, form, puget, service, marketplace, base, team, product, internet

Glossary: eps: earnings per share; puget sound energy company

## Topics from plot (2,2)

200 docs [trained model n460], BIC $\approx -104898$, $n = 61$

Topic A: estimate, buy, expect, strong, market, point, percentage, perform, eps, rate, usd, reduce, underperform, outperform, continue

Topic B: billion, enron, credit, stock, financial, commercial, paper, investor, bank, line, trade, corp, chief, company, fall

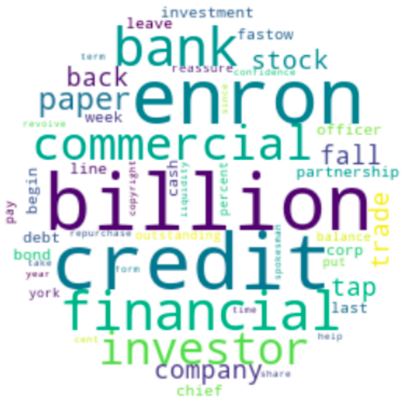
Topic C: gas, power, natural, customer, energy, electric, co, houston, plan, electricity, bill, texas, consumer, state, newpower

Topic D: brother, lehman, expect, estimate, inc, point, market, strong, percentage, director, buy, underperform, reduce, ercot, perform

Glossary: ercot: electric reliability council of texas

## Possible fraud?

180 docs [trained model n410], BIC $\approx -91151$, $n = 64$



billion, credit, enron, financial, bank, commercial, investor, paper, company, stock, trade, fall, tap, back, partnership,
debt, chief, investment, corp, officer, cash, line, bond, fastow, week, leave, begin, last, outstanding, pay, ... , balance, ... , repurchase, ... , liquidity, ...

## Remark

- Independent of the test BIC general behaviour (though favour a decreasing manner)

- Independent of the lower test BIC value exhibited across the groups

- The interpretable topics shown persist as well under other models. (There are a few other topics noted.)

- The documents grouped form a (much smaller) subset of the entire data. Given the certain degree of interpretability, perhaps one can dive in the documents if a cloud gives itself away.