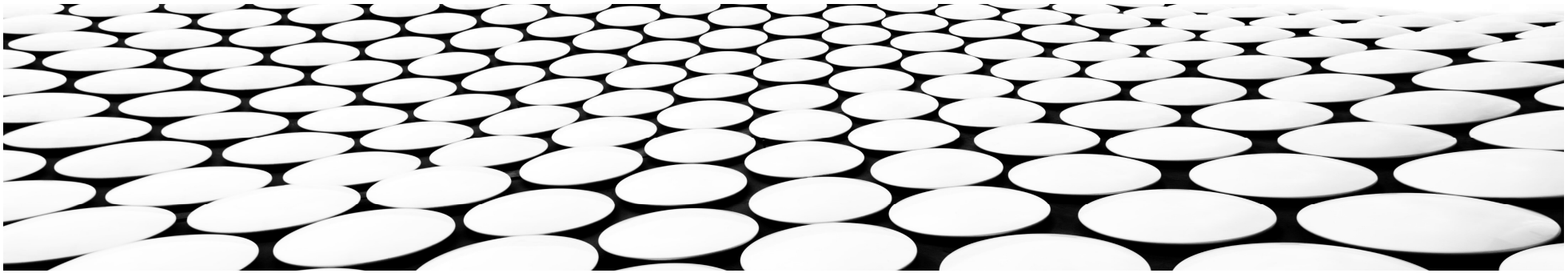
TOPIC MODELING : DEEP DIVE INTO TEXT ANALYTICS

SANIL MHATRE

LEAD DATA SCIENTIST, WORD WIDE TECHNOLOGY



SANIL MHATRE

LEAD DATA SCIENTIST
WORLD WIDE TECHNOLOGY

/SanilMhatre

@sqlsuperguru

- Data Scientist with extensive background in Data Engineering, Business Intelligence, Database Administration and Enterprise Architecture
- Azure, AWS, GCP
- SQL Sever, Oracle, Snowflake, PostgreSQL, MongoDB
- Agile Coach, Mentor, Speaker, Blogger, Volunteer

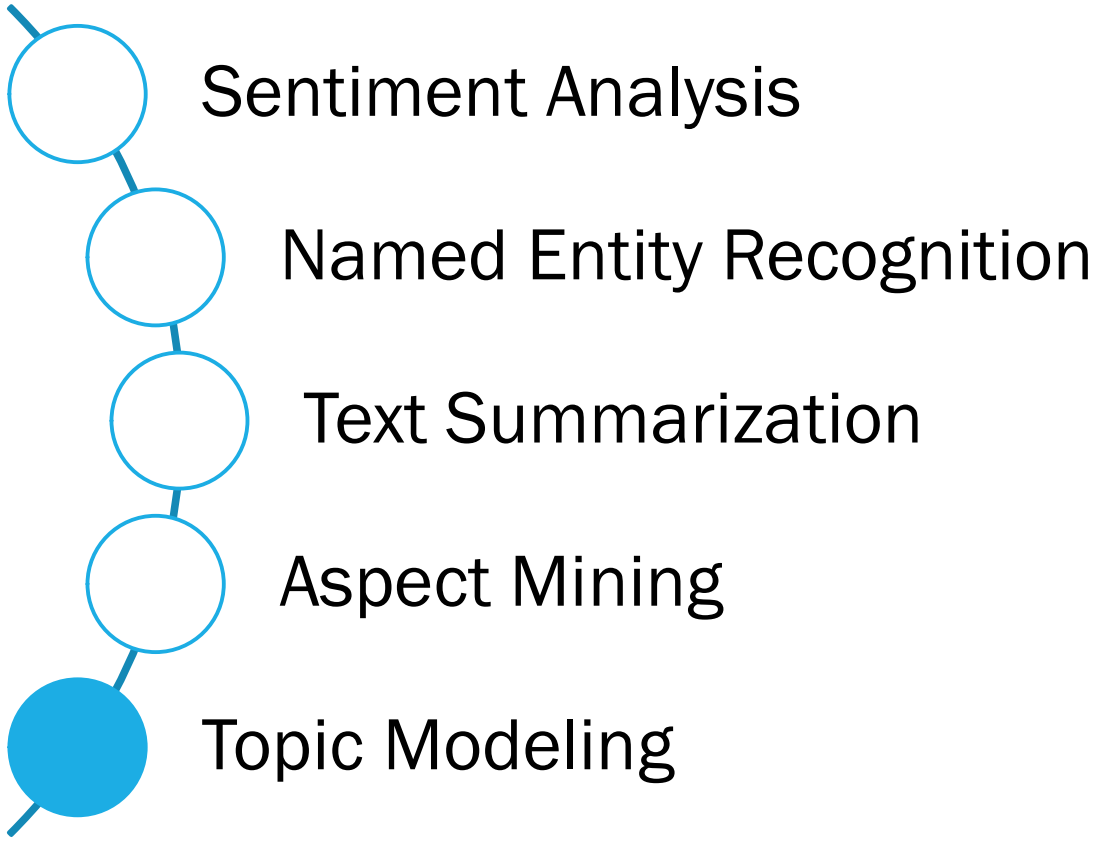
https://www.red-gate.com/simple-talk/author/sanil-mhatre/

AGENDA

- Topic Modeling
- Data Prep
- Model Training & Evaluation
- Topic Inference

NATURAL LANGUAGE PROCESSING

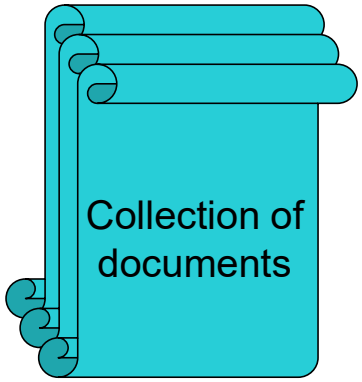
Machine Learning driven process of deciphering human language text data using software



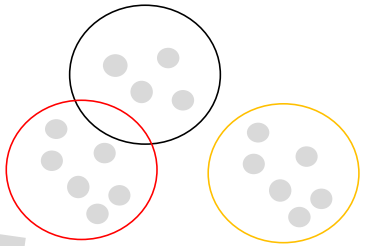
TOPIC MODELING

- Unsupervised Machine Learning technique
- Represents text document as collection of topics
- Finds relationships amongst data in text documents
- Latent Dirichlet Allocation (LDA) probabilistic modeling
  - Text document as distribution of Topics
  - Topics as collection of words

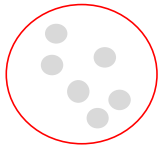
LATENT DIRICHLET ALLOCATION



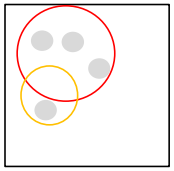
LDA Algorithm

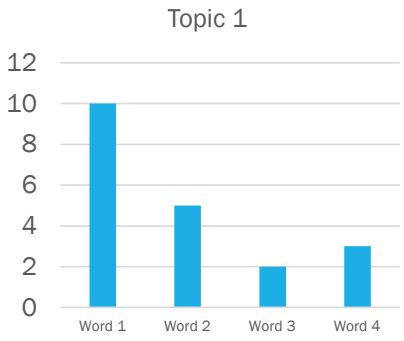


3 topics identified – each topic is a cluster of words



Each document is a collection of topics – *Topics are distributed across each document*

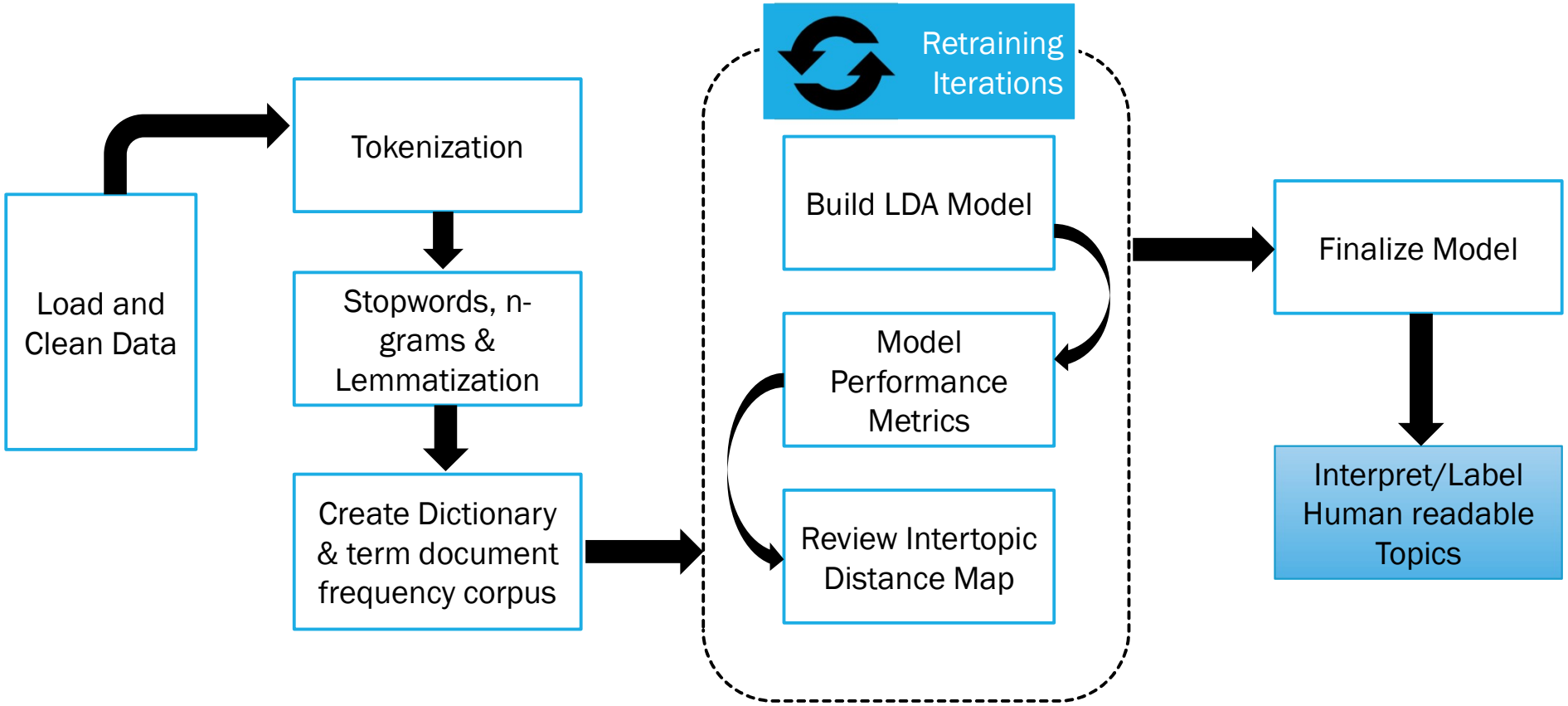




WHY USE TOPIC MODELING ?

- Need for meaning and actionable insights
- Limits of Text summarization, Aspect mining and key word/phrase frequency techniques
- Topic modeling is relatively easy, reliable and popular technique
- Discover valuable business insights with topic modeling
  - Top 5 customer complaints from on chat text/call transcripts
  - Top 3 suggestions for improvements from survey text

TOPIC MODELING PROCESS



DEMO USE CASE

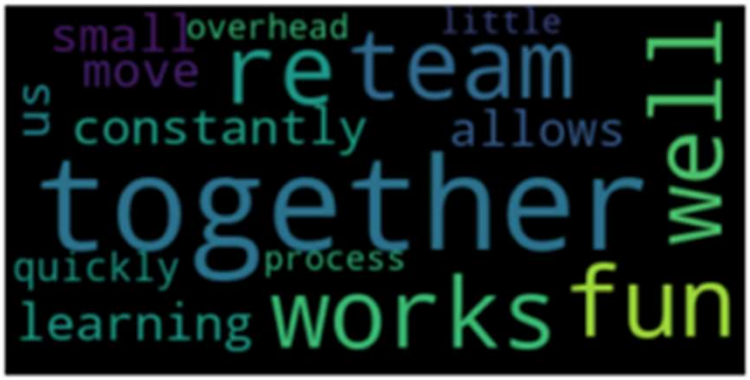
- Quarterly Survey of 9 IT Teams
- One open ended question : How do you feel about your team's health in this Quarter ?
- 9 Teams, 3 Managers, 4 Quarters
- 300 Responses
- Raw Data in Excel

DEMO ENVIRONMENT

- Anaconda, Jupyter Notebook, Python
- spaCy – Lemmatization
- NLTK – Stopwords
- Gensim – Topic Modeling
- pyLDAvis – interactive web-based visualization
- Additional Libraries/Packages
  - Pandas, NumPy, Matplotlib, Wordcloud

DEMO : WORD CLOUD

- Load data from Excel file
- Simple data cleaning steps
- Generate Word Cloud
- Interpret Word Cloud



DEMO : TOKENIZATION & STOP WORDS

- Tokenization is the process of separating a body of text into smaller units called “tokens”, to apply NLP techniques
  - Tokens can be words, phrases (n-grams) or characters
- Stop Words don’t add much value to a sentence and can be ignored without compromising it’s meaning
  - Stop words are filtered out before further processing
  - Pre-defined stop list for most languages built into popular packages
  - Popular packages allows customization/extension of stop list

DEMO : N-GRAMS & LEMMATIZATION

- N-Grams
  - Bigrams – sentence/phrase composed of two words
  - Trigrams - sentence/phrase composed of three words
  - Examples : “pretty good” , “lack of”
- Lemmatization
  - Process of converting words to their roots
  - Uses contextual vocabulary and morphological analysis
  - More effective than stemming
  - Example : “Walk” is the lemma (root) of “walking” & “walks”

DEMO : DICTIONARY & CORPUS

- LDA topic model needs “dictionary” and “corpus” as inputs
- Dictionary
  - Collection of lemmatized words from the text
  - Unique id assigned to each word
- Corpus
  - Latin for “body”, refers to a collection of texts
  - Term document frequency corpus
  - Uses “unique id” from dictionary

DEMO : BUILD LDA TOPIC MODEL

- Gensim library module : Gensim.models.ldamodel
- Key Inputs
  - “corpus” & “dictionary” created previously
  - “num_topics” (iterate - 2 to …….. n)
- Output
  - Topic id
  - key words for each Topic
  - Importance Score for each keyword

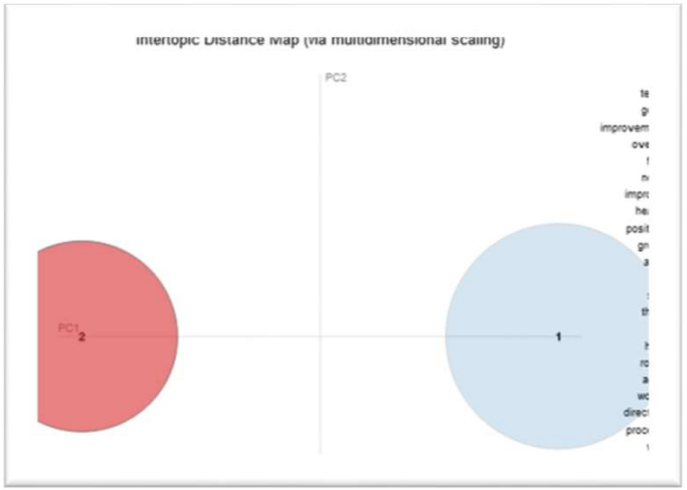
DEMO : MODEL PERFORMANCE METRICS

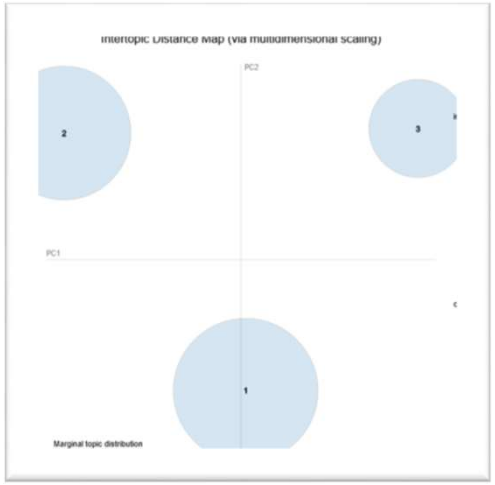
- Perplexity
  - Measure of how “surprised” a model with new data
  - Normalized log-likelihood of a held-out test set
  - Lower value is better
- Topic Coherence
  - Set of facts/statements are “coherent” if they support each other
  - Topic coherence measure semantic similarity between words of same topic
  - Higher value is better

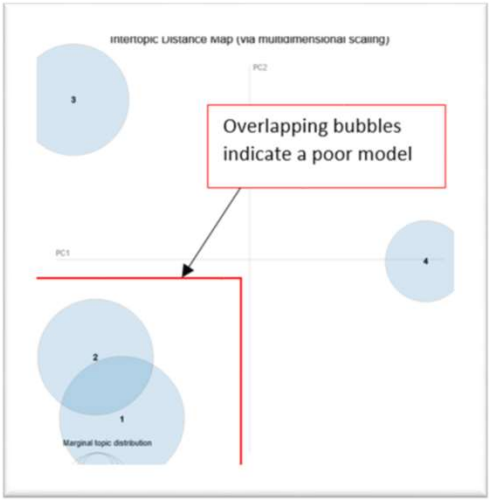
DEMO : INTERTOPIC DISTANCE MAP

- Optimizing for Perplexity & Topic Coherence may not always lead to human interpretable topics
- pyLDAvis package - Intertopic Distance map
  - Interactive web-based visualization
  - Each bubble represents a topic
  - Size of bubble represents its prevalence
  - Large, non-overlapping & scattered bubbles are optimal

DEMO : OPTIMAL NUMBER OF TOPICS

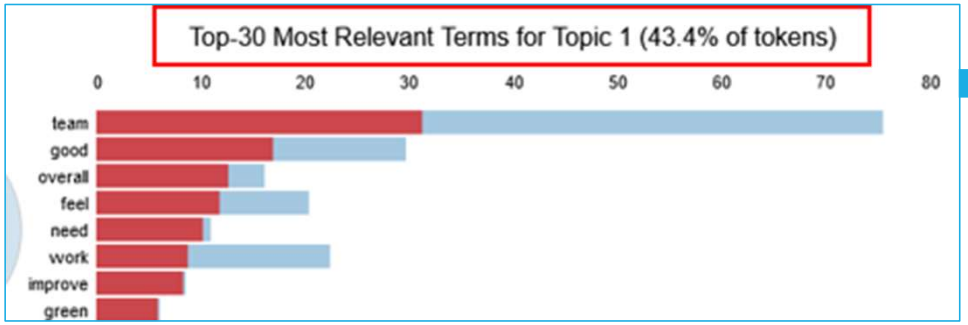




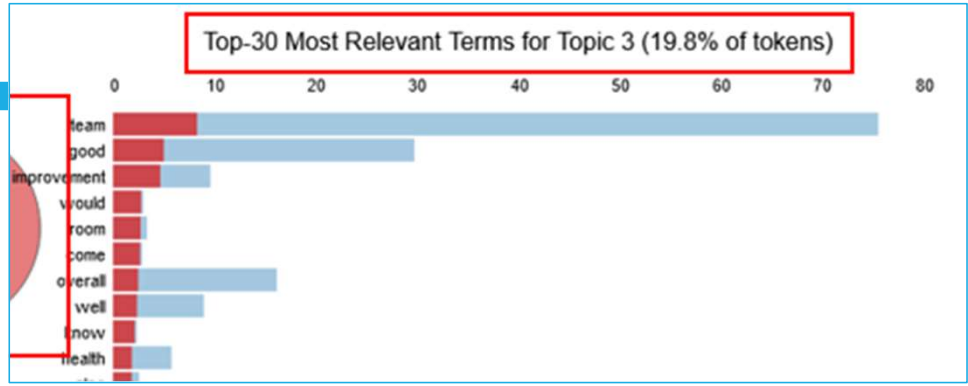


Num_topics	Model perplexity	Topic Coherence	Intertopic Distance Map
2	-6.089	0.221	Two large bubbles well-spaced across chart quadrants
3	-6.174	0.245	Three large bubbles well-spaced across chart quadrants
4	-6.253	0.274	Three large bubbles and one small. Bubbles for topics 1 and 2 are overlapping

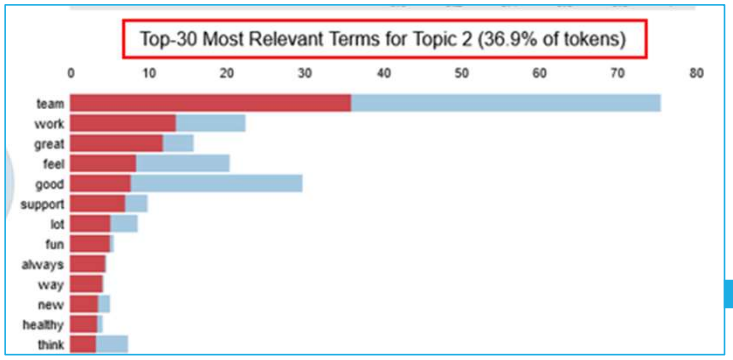
DEMO : INFER TOPIC LABELS



Top 10 terms of topic 1 are used to infer label "Overall Team (health) feel(s) good, positive (and) green"



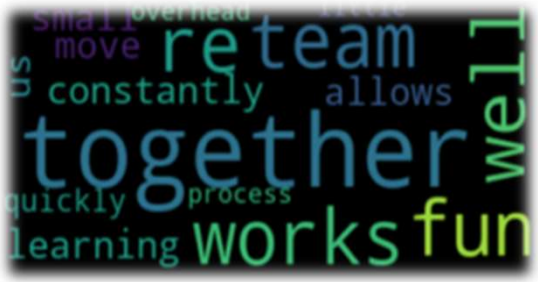
Top 10 terms of topic 2 are used to infer label "Great work, good support (and) lot (of) fun"



Top 10 terms of topic 3 are used to infer label "Room (for) improvement"

DEMO : COMPARE AGAINST WORD CLOUD

WORD CLOUD



- visual indicates “together” is the most frequent key word in this text, followed by “works”, “fun”, “well” and “team”
- While these words have positive connotations, it’s difficult to gain deeper insights from this word cloud

TOPIC MOELING

Topic Number	Percentage Composition of Tokens	Topic Label
1	43.3 %	Overall Team health is good/positive
2	36.9 %	Great work, lot of fun and supportive team
3	19.8 %	Some room for improvement

Combined with business context, subject matter expertise and organizational knowledge, the following might be a good a summary readout to potential business stakeholder group of IT leadership.

- The prevalent consensus (43.3 %) amongst survey respondents indicates that overall Team health is good/positive
- Over a third (36.9 %) of survey responses indicate teams are supportive of their members, they have lot of fun and work is great (positive environment for teamwork)
- Around one fifth (19.8 %) of survey responses hint at some room for improvement

AUTOMATION

- Program a Loop
  - Write a loop to iterate the num_topics from “2” to “30”
  - Plot model performance metrics
  - Plot pyLDAvis chart for each iteration and review the Intertopic Distance Maps to find the optimal number of human readable topics
- LDA Mallet Model
  - Mallet is an open-source toolkit for NLP with a package for LDA based topic modeling
  - Gensim provides a wrapper to facilitate Mallet’s LDA topic model estimation and inference of topic distribution

CONCLUSION

- Introduce NLP technique of Topic Modeling
- Setup of Anaconda Jupyter notebook environment for performing topic modeling
- Data cleaning and preparation steps needed for topic modeling with LDA
- Iterative process of training topic models and identifying an optimal solution
- Interpreting human readable insights from topic model output charts
- Comparing these deeper insights with outcomes from the easier technique of word cloud
- Business value of topic modeling as a popular and practical Natural Language processing technique

REFERENCES

- Topic modeling - https://en.wikipedia.org/wiki/Topic_model
- Latent Dirichlet allocation - https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation
- LDA paper from Journal of Machine Learning Research - https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
- TF-IDF - https://en.wikipedia.org/wiki/Tf%E2%80%93idf
- Lemmatization and stemming - https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html
- spaCy - https://spacy.io/
- tokenization - https://aclanthology.org/C92-4173.pdf
- n-grams - https://en.wikipedia.org/wiki/N-gram
- Evaluate topic models using perplexity and coherence scores - https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0
- pyLDAvis - https://pyldavis.readthedocs.io/en/latest/readme.html