Oppdag det forventede og uventede ved å bruke tekstanalyse i VA

FANS Nettverksmøte – 22. og 23.mai 2024

Vegard Hansen
Academic Lead @ SAS
Global Academic Program, Education
vegard.hansen@sas.com
www.linkedin.com/in/vegard-hansen/

Copyright © SAS Institute Inc. All rights reserved.

SAS Global Academic Program

Creating a Pipeline of Skilled Future Users

- Build SAS Skill Aligned with Workforce Demand
- Provide Free Teaching and Learning Resources & Platforms and Recognition
- Connect Graduates with SAS Customers for Hiring

SAS (Certified) Specialist

SAS SKILL BUILDER for Students

Self-study portal for students

Totally free

From Programming to Machine Learning

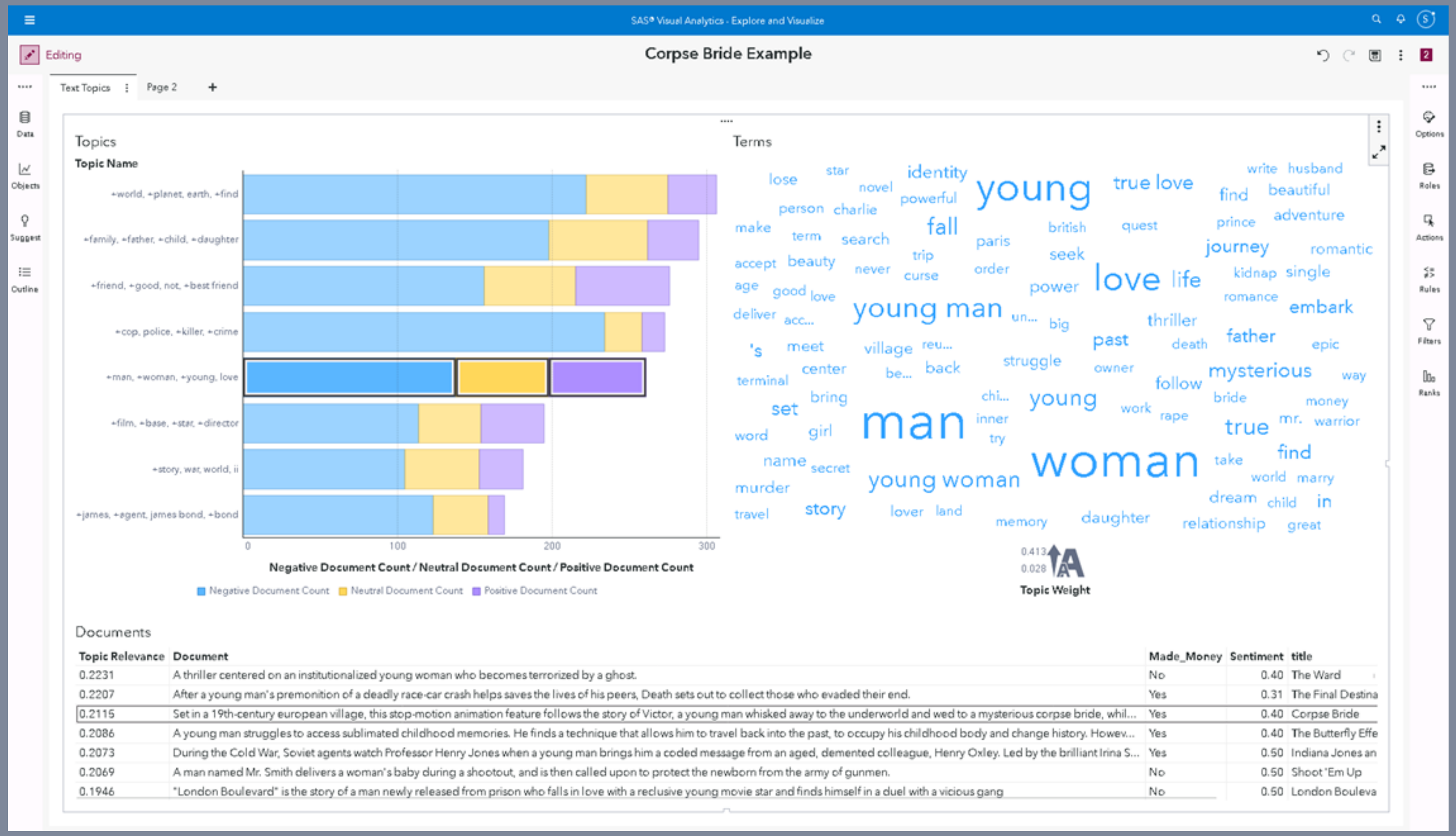
The + of Badges on the CV

SAS EDUCATORS Portal

Teacher Resource Portal

Totally free

How text/unstructured data can be used to improve the value of your predictive models



Text Topics Object in SAS VA

Copyright © SAS Institute Inc. All rights reserved.

Text or unstructured data can significantly improve the value of predictive models

1. **Sentiment Analysis**:

   Analyses sentiments from text data like reviews or comments to understand user preferences and trends, which can be used to make more accurate predictions.

2. **Topic Modeling**:

   Identifies topics within the text data to uncover hidden patterns and insights that can be useful for prediction.

3. **Named entity recognition (NER)**

   NER is a text analytics technique used for identifying named entities like people, places, organizations, and events in unstructured text.

4. **Term frequency – inverse document frequency**

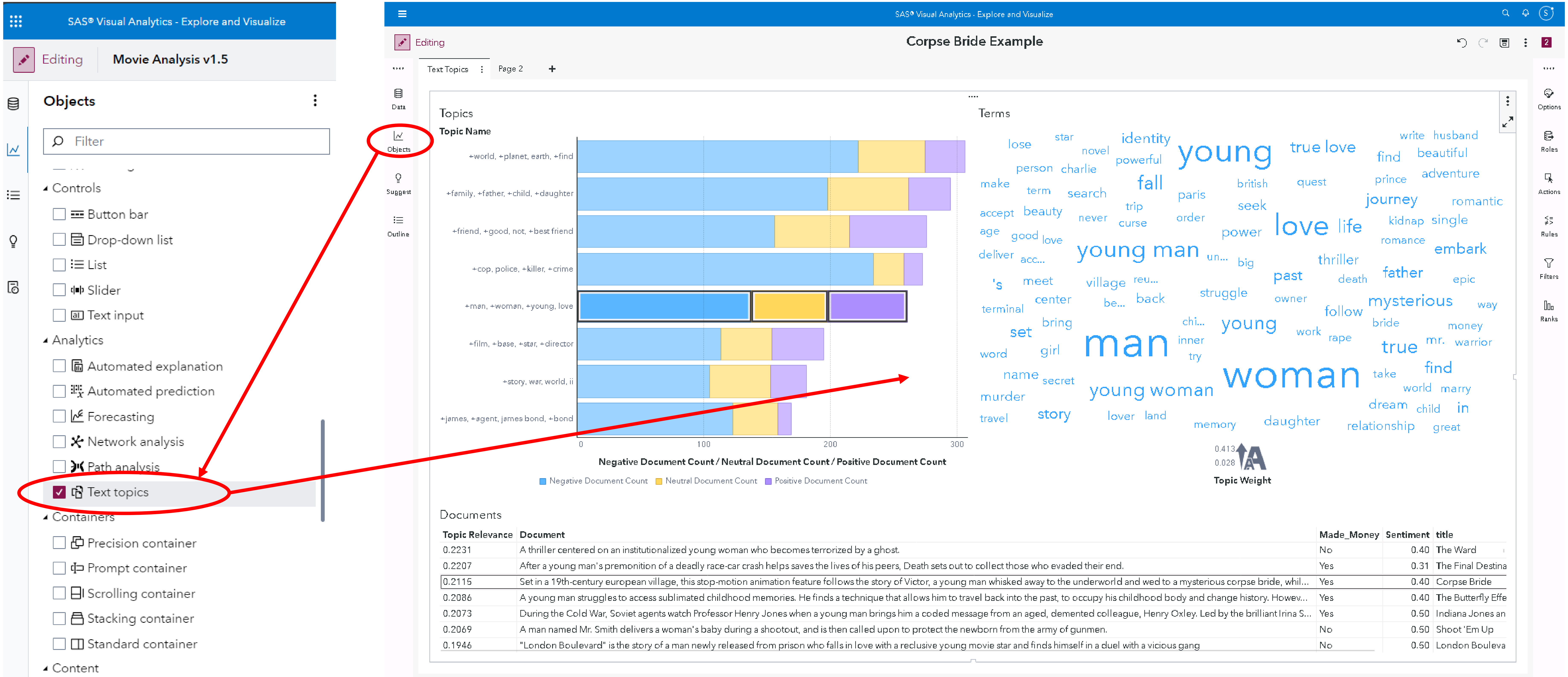
   TF-IDF is used to determine how often a term appears in a large text or group of documents and therefore that term's importance to the document.

5. **Event extraction**

   This is a text analytics technique that is an advancement over the named entity extraction. Event extraction recognizes events mentioned in text content, for example, mergers, acquisitions, political moves, or important meetings.

Copyright © SAS Institute Inc. All rights reserved.

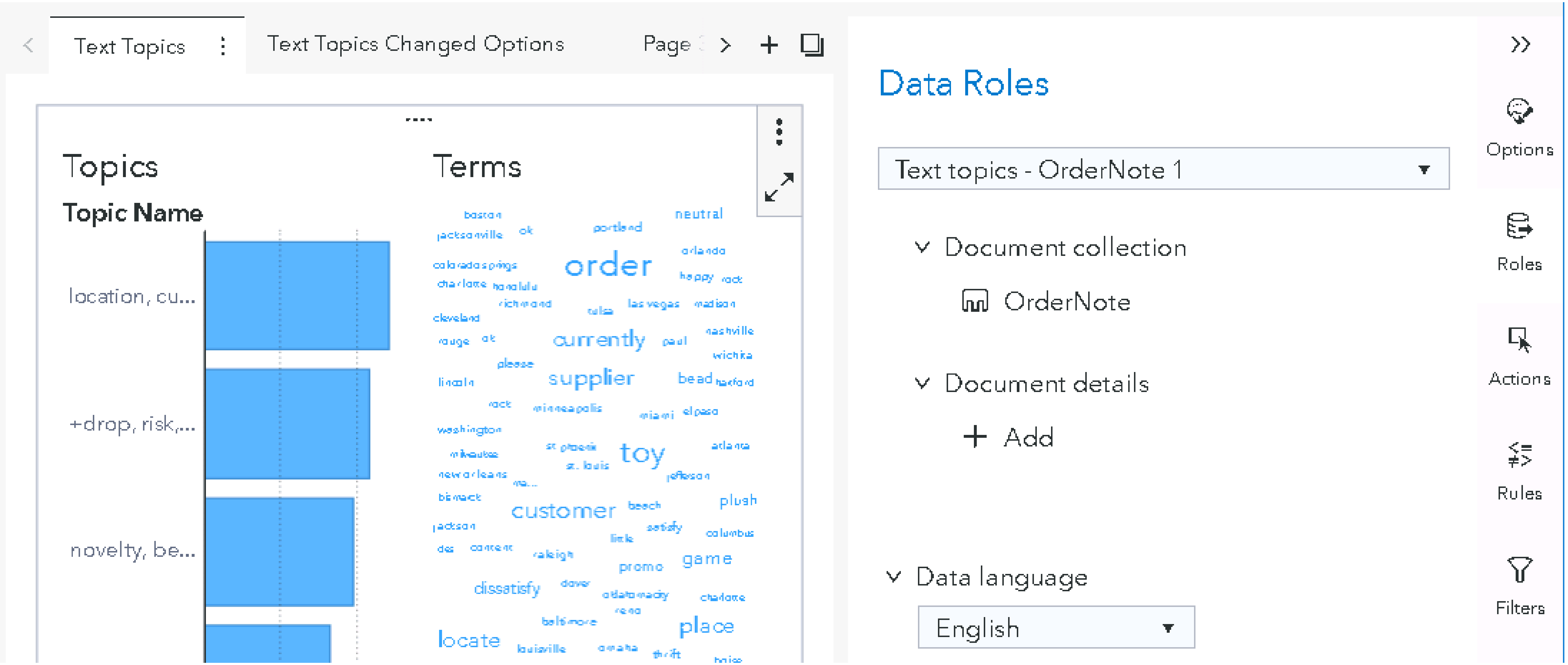
Meet the Text Topics Object



Copyright © SAS Institute Inc. All rights reserved.

Data Roles

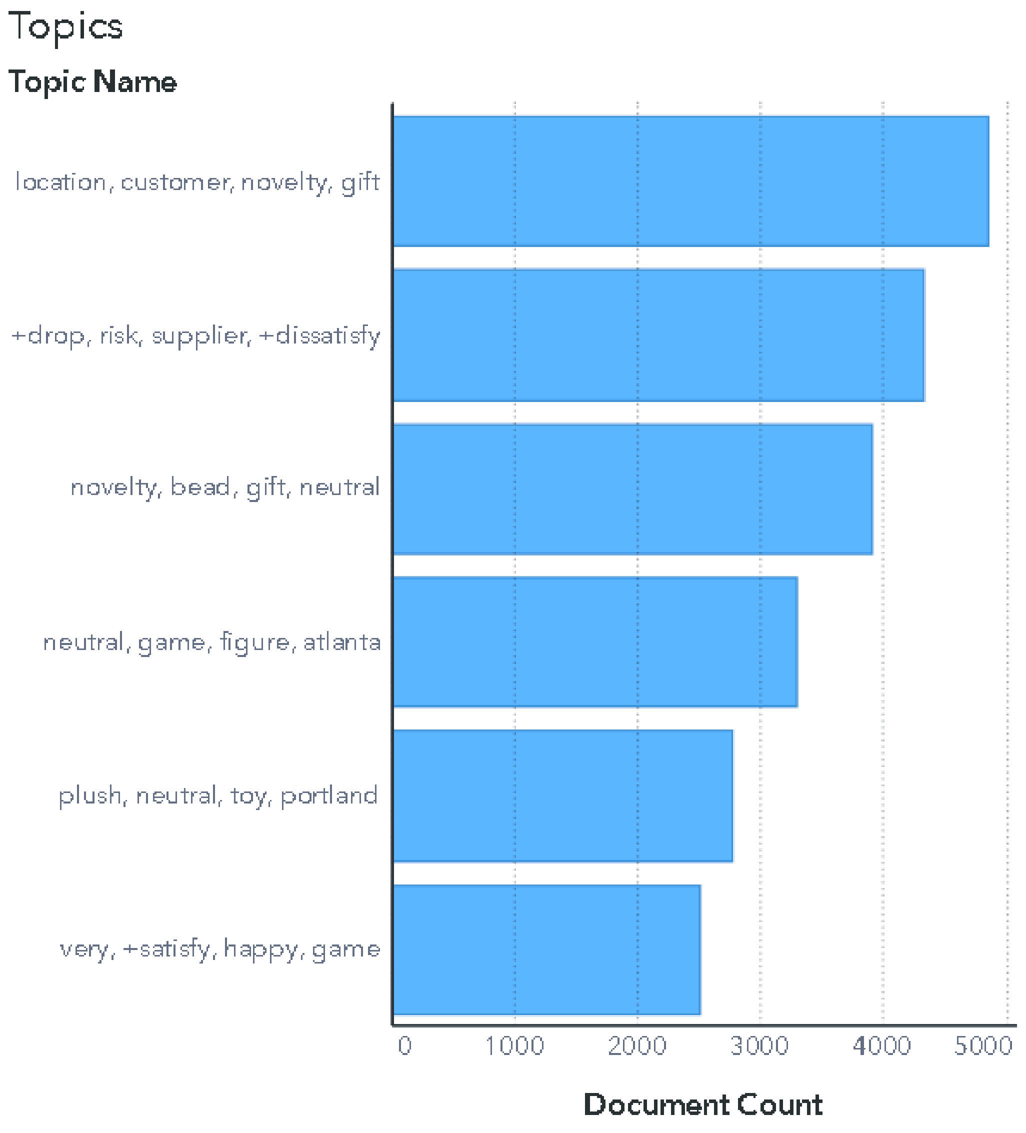
There must be a unique ID in your data!



Copyright © SAS Institute Inc. All rights reserved.

Output

- Suggested topics
- Document count for each topic
- Word Cloud with word frequency

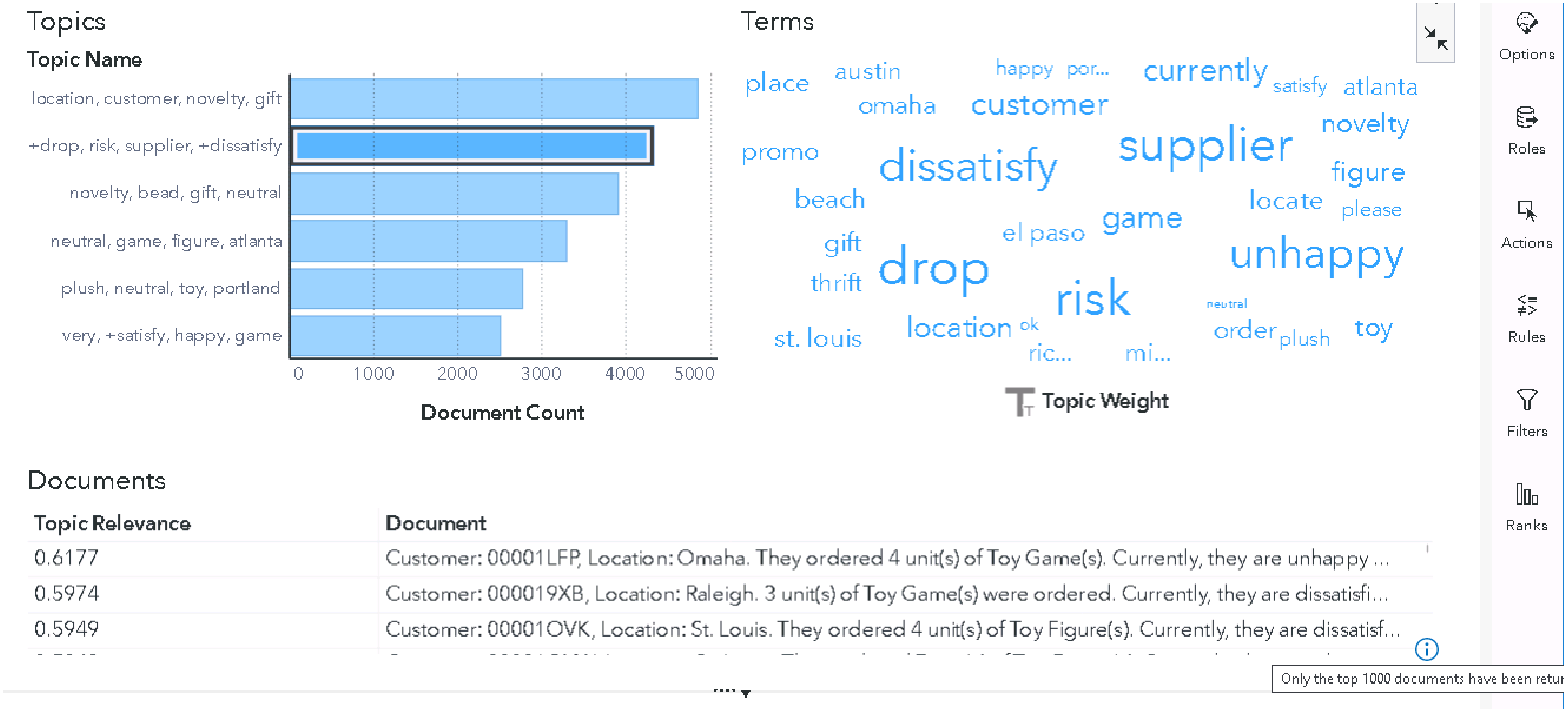


Copyright © SAS Institute Inc. All rights reserved.

Selected Topic

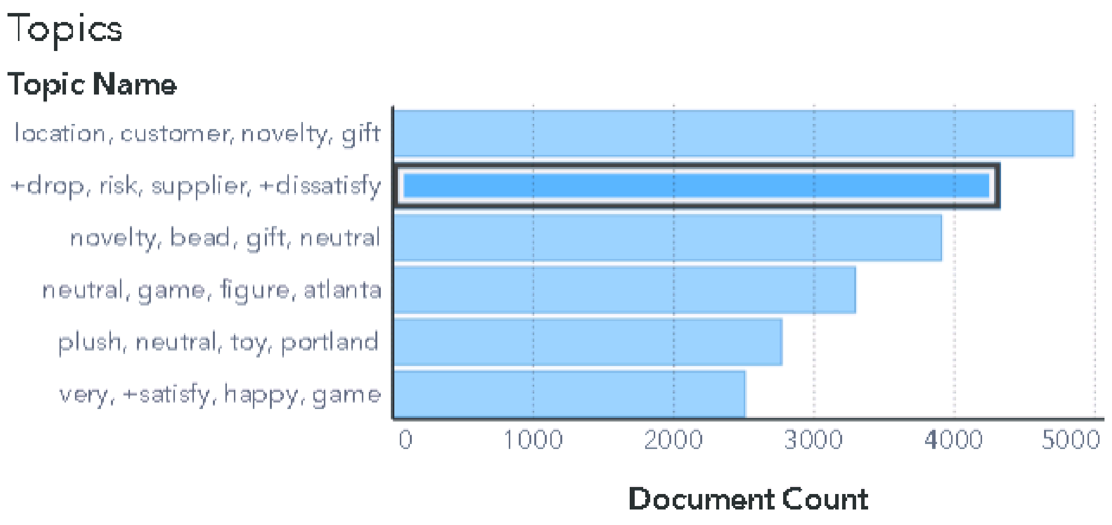
Click on a topic

- Word Cloud shows term weights related to the selected topic
- Documents are shown ranked by topic relevance



Copyright © SAS Institute Inc. All rights reserved.

Details Table



Documents

Topic Relevance	Document
0.6177	Customer: 00001LFP, Location: Omaha. They ordered 4 unit(s) of Toy Game(s). Currently, they are unhappy with us as ...
0.5974	Customer: 000019XB, Location: Raleigh. 3 unit(s) of Toy Game(s) were ordered. Currently, they are dissatisfied with u...
0.5949	Customer: 00001OVK, Location: St. Louis. They ordered 4 unit(s) of Toy Figure(s). Currently, they are dissatisfied with ...

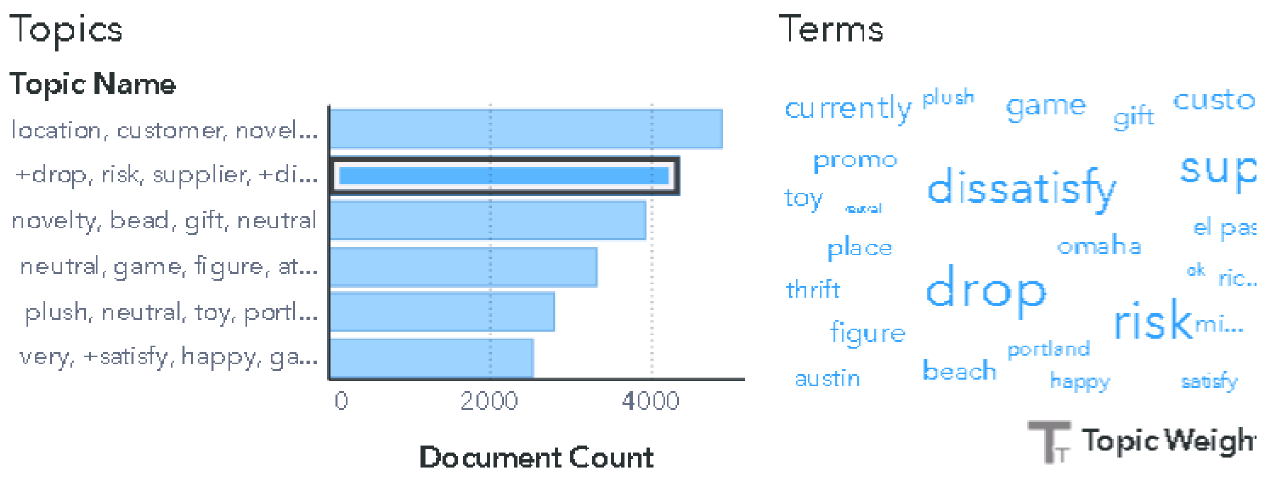
Topics | Terms | Text Topics Summary

Topic Name	Document Count
location, customer, novelty, gift	4859
+drop, risk, supplier, +dissatisfy	4331
novelty, bead, gift, neutral	3913
neutral, game, figure, atlanta	3301
plush, neutral, toy, portland	2777
very, +satisfy, happy, game	2514

Copyright © SAS Institute Inc. All rights reserved.

Terms Table

is a numerical representation of the Word Cloud



Documents

Topic Relevance	Document
0.6177	Customer: 00001LFP, Location: Omaha. They ordered 4 unit(s) of
0.5974	Customer: 000019XB, Location: Raleigh. 3 unit(s) of Toy Game(s)
0.5949	Customer: 00001OVK, Location: St. Louis. They ordered 4 unit(s)

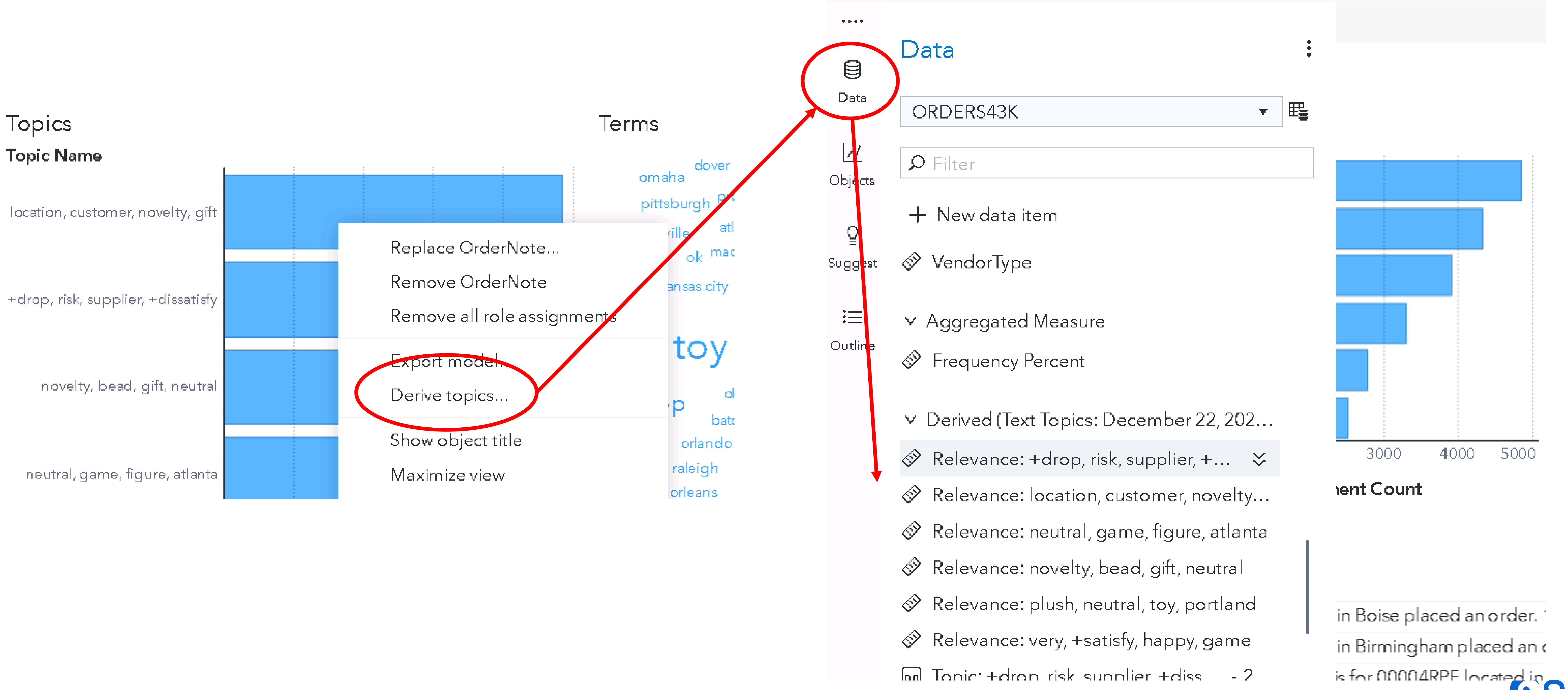
Topics | Terms | Text Topics Summary

Term	Topic Weight	Role
drop	0.475	Verb
risk	0.438	Noun
supplier	0.408	Noun
dissatisfy	0.356	Verb
unhappy	0.347	Adjective
customer	0.156	Noun
game	0.139	Proper noun
currently	0.138	Adverb
location	0.106	Noun

Copyright © SAS Institute Inc. All rights reserved.

Derive Topics

with a right mouse click

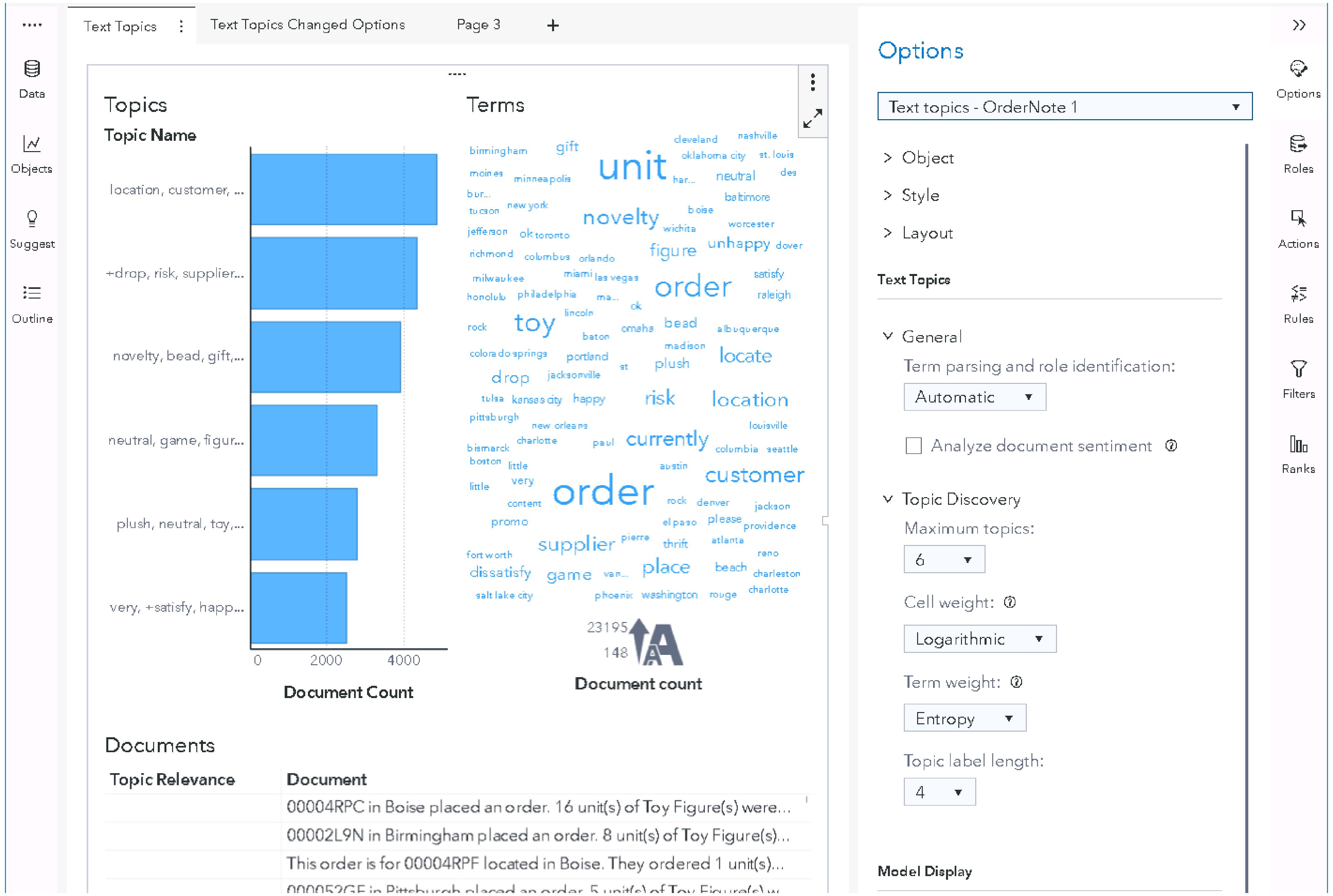


Copyright © SAS Institute Inc. All rights reserved.

Options for Text Topics

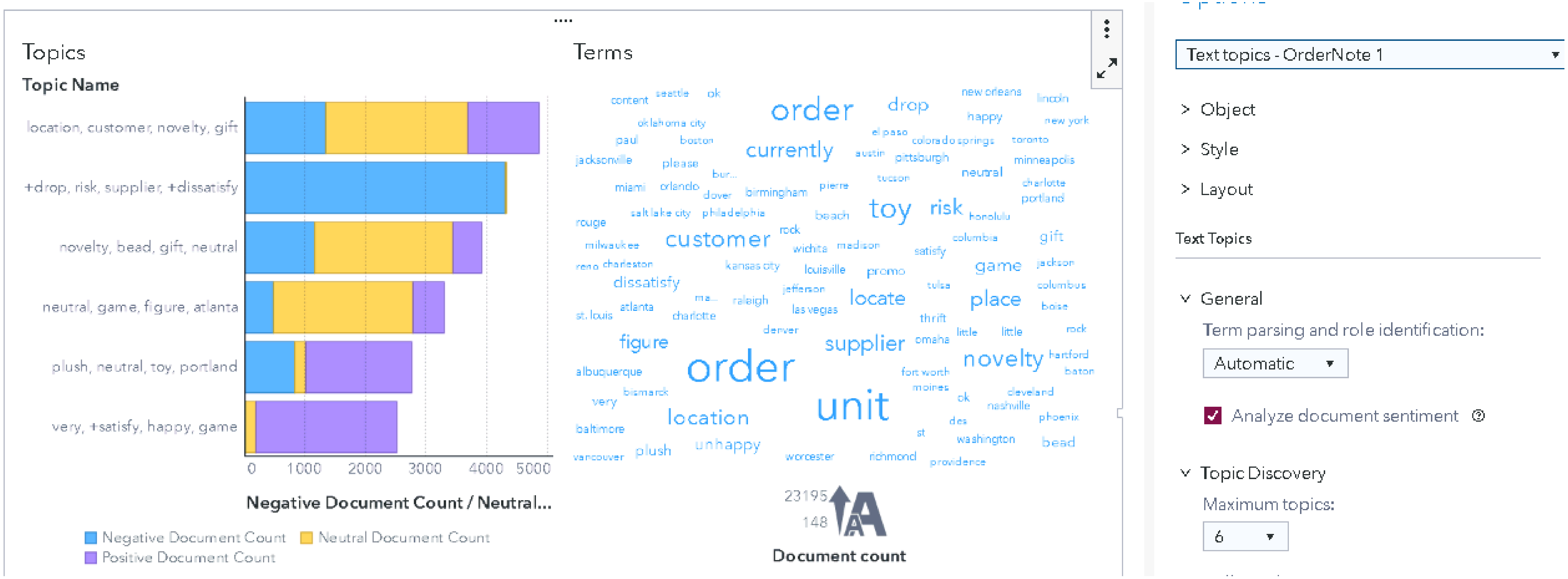
You can adjust:

- Parsing (how the terms are built)
- The maximum number of topics
- Emphasize rare terms
- The length of the topic name
- Sentiment Analysis



Copyright © SAS Institute Inc. All rights reserved.

And if you do choose to analyze sentiments…



Copyright © SAS Institute Inc. All rights reserved.

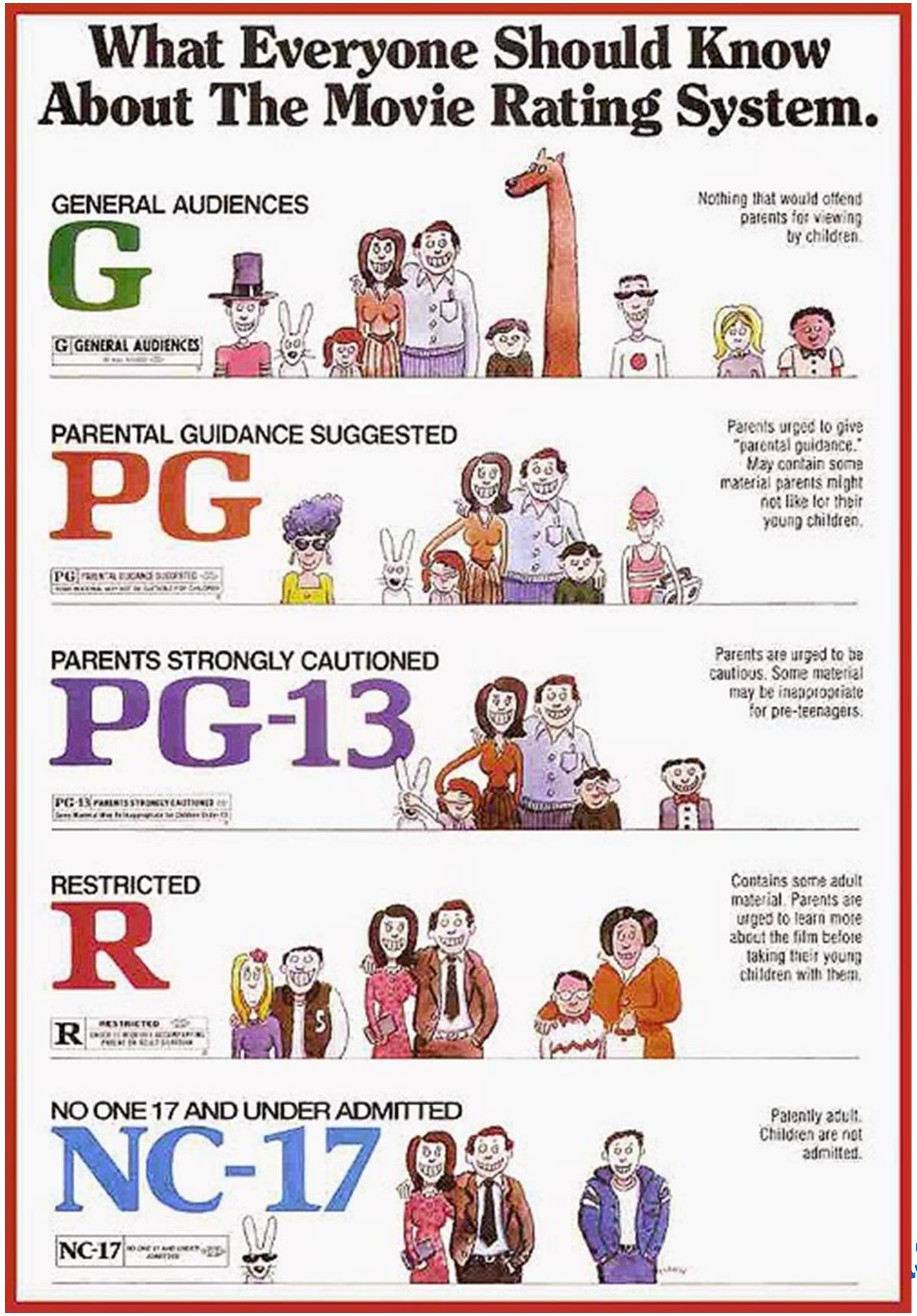
- What: A boutique firm specializing in movie consulting
- Specifically: Advise movie production companies on which new movie proposal could be a potential smash hit.
- Analytical task/goal: Based on analysing the characteristics of historical films find what type of movies that maximise Viewer Rating scores.
- Data: dataset of 1500 historical movies with viewer rating (score 1-5) and synopsis (unstructured text data) together with other data

Copyright © SAS Institute Inc. All rights reserved.

Data - Moviedata

Variables

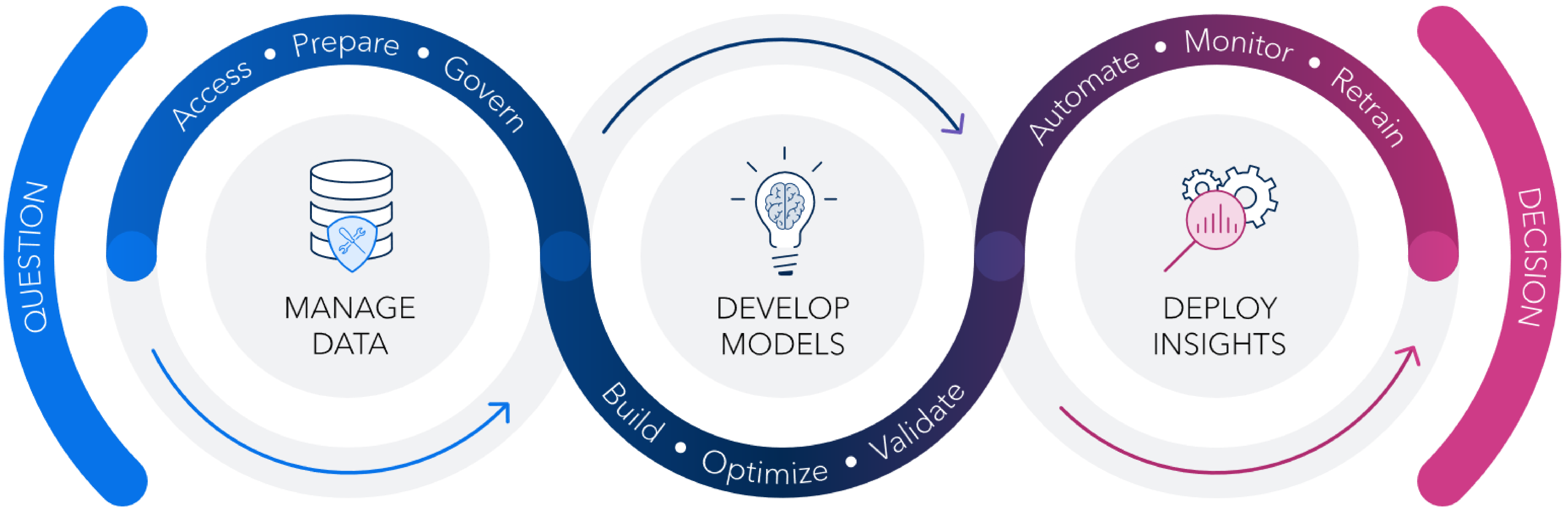
- Synopsis – overview text
- Title – unique
- MPAARating – Rating (R,PG,PG-13 NR,G,NC-17)
- Genre
- Year
- **Viewer Rating – Target (0-4)**
- Size
- … (Dummy ++)



Copyright © SAS Institute Inc. All rights reserved.

Operationalizing Analytics

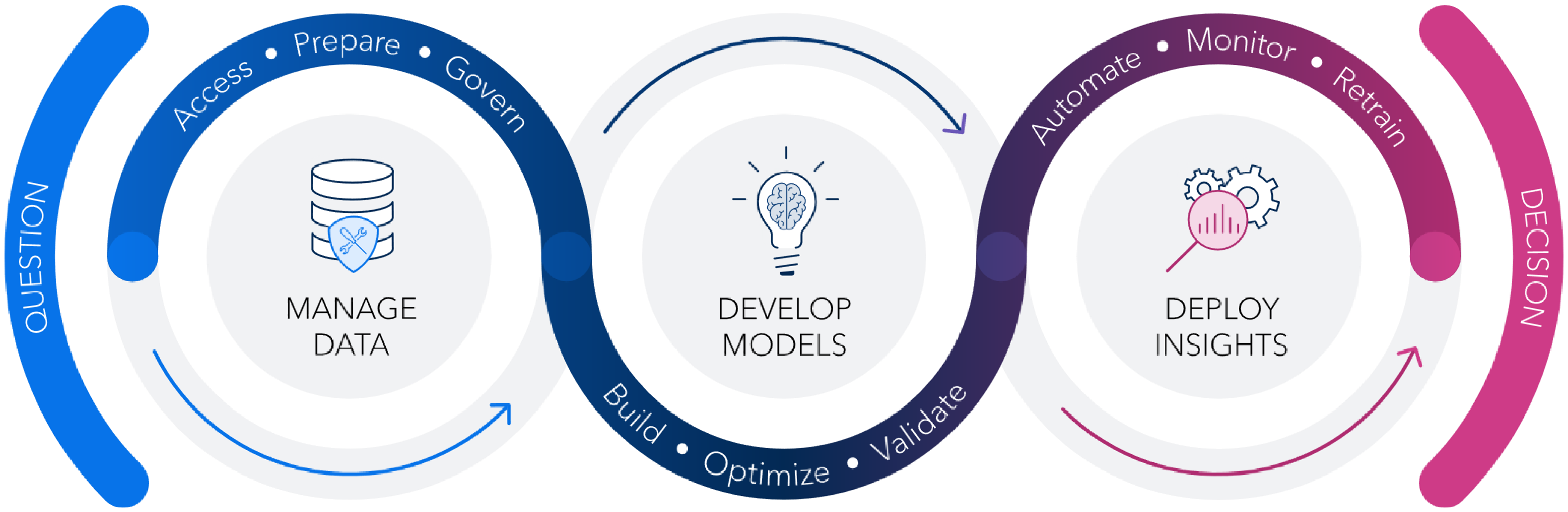
The Analytics Lifecycle



Copyright © SAS Institute Inc. All rights reserved.

Operationalizing Analytics

Which new movie proposal could be a potential smash hit by analysing the characteristics of historical films?



Decide which type of movie that maximizes the Viewer Rating

- Step 1: Data Access: MovieGenres - structured + unstructured data
- Step 1: Understand your data:
    - Data preparation and exploration
    - Define your target – Viewer Rating
    - Initially analytics and explorations
- Step 2: Add text prepossessing:
    - Create topics for modelling

- Step 1: Add a regression model to your (structured) data, measure models performance: ASE
- Step 2: Based on outcome of the text analysis extend your regression model in step 1 with “new” features
- Compare the two models and choose the best one (champion)

Finally, you will be ready to;

- Bring your champion model to further analysis and deployment
- Deploy the results in a Dashboard
- Validate samples (type of movies) with experts - ASE
- In long Run:
    - Score the model against new data
    - Monitor periodically (quarterly, yearly, when needed…)

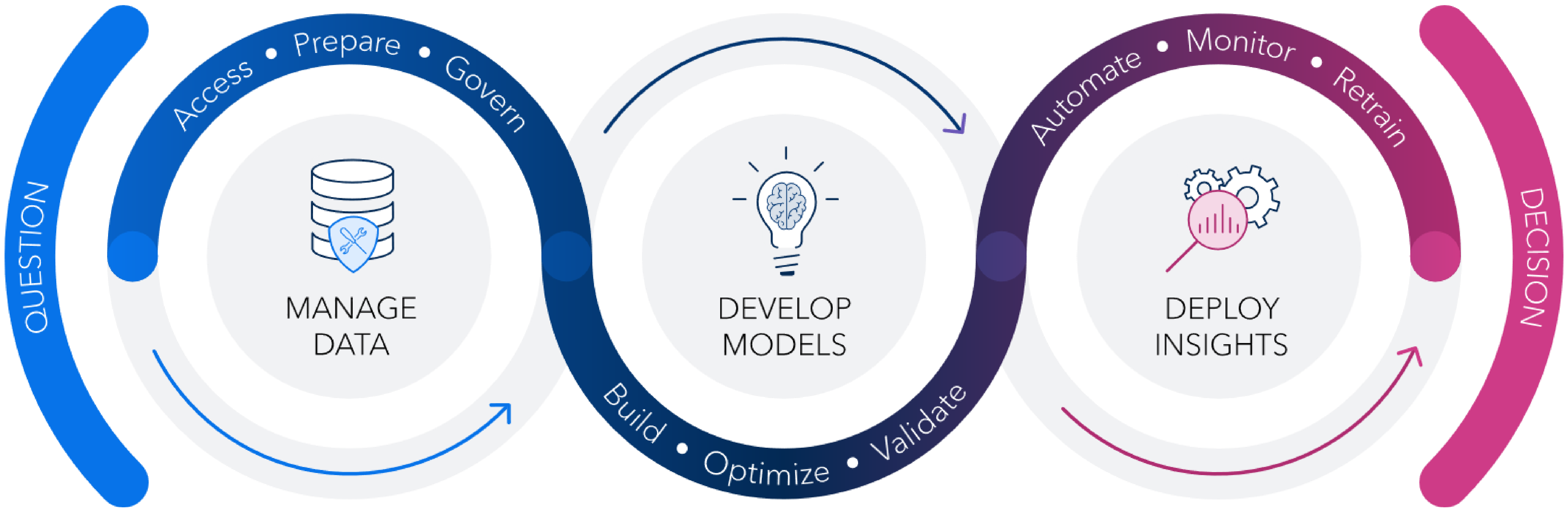
Copyright © SAS Institute Inc. All rights reserved.

Sentiment Example: Customer review sentiment classifier

How can texted review data improve a predictive model?

Review: *"I absolutely love this product! It's fantastic."*
Sentiment Label: *<u>Positive</u>*

Decide whether a given review is positive or negative based on the text content.



1. Data Collection – Custom reviews – text + sentiment (target)
2. Understand and know your data, i.e.:
    - Data preparation and exploration
    - Analyse and find (hidden) insight
3. Text preprossesing:
    - Tokenzise, start/stop, stemming etc….
    - Feature Extraction

Model Building:
- Train a (ML) model using the feature vectors
- The model learns to associate certain word patterns with positive or negative sentiments
- Validate: Assess the model's performance Sentiment Prediction:
- Compare models and choose champion

- Deploy the model to the e-commerce platform
- Score new reviews – when customer submits a new review
    - The model predicts whether the review expresses a positive or negative sentiment
- Monitor – performance, retrain, retire, replace….
    - Using metrics like accuracy, precision, recall, or F1-score. Fine-tune hyperparameters if necessary.
    - Preprocess and convert into a feature vector.
    - Feed the vector into the (new) model.

Copyright © SAS Institute Inc. All rights reserved.

Spørsmål?

#learnsas

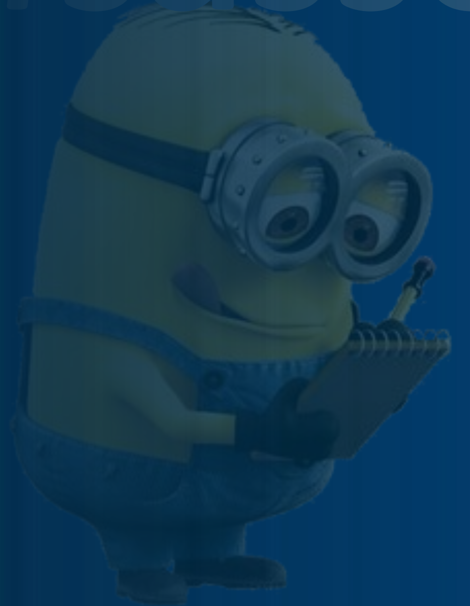
#skillbuilder

#lifelonglearner

#securethefuture

#sassoftware

#weareallacademics



Vegard Hansen
Academic Lead @ SAS
vegard.hansen@sas.com
www.linkedin.com/in/vegard-hansen/

SAS®

Copyright © SAS Institute Inc. All rights reserved.