

Oppdag det forventede og uventede ved å bruke tekstanalyse i VA

FANS Nettverksmøte – Onsdag 13.mars 2023

fans*

Vegard Hansen

Academic Lead @ SAS

Global Academic Program, Education

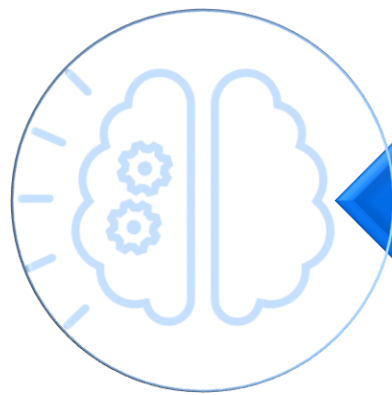
vegard.hansen@sas.com

www.linkedin.com/in/vegard-hansen/



SAS Global Academic Program

Creating a Pipeline of Skilled Future Users



Build SAS Skill Aligned with Workforce Demand



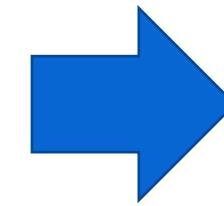
Provide Free Teaching and Learning Resources & Platforms and Recognition



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

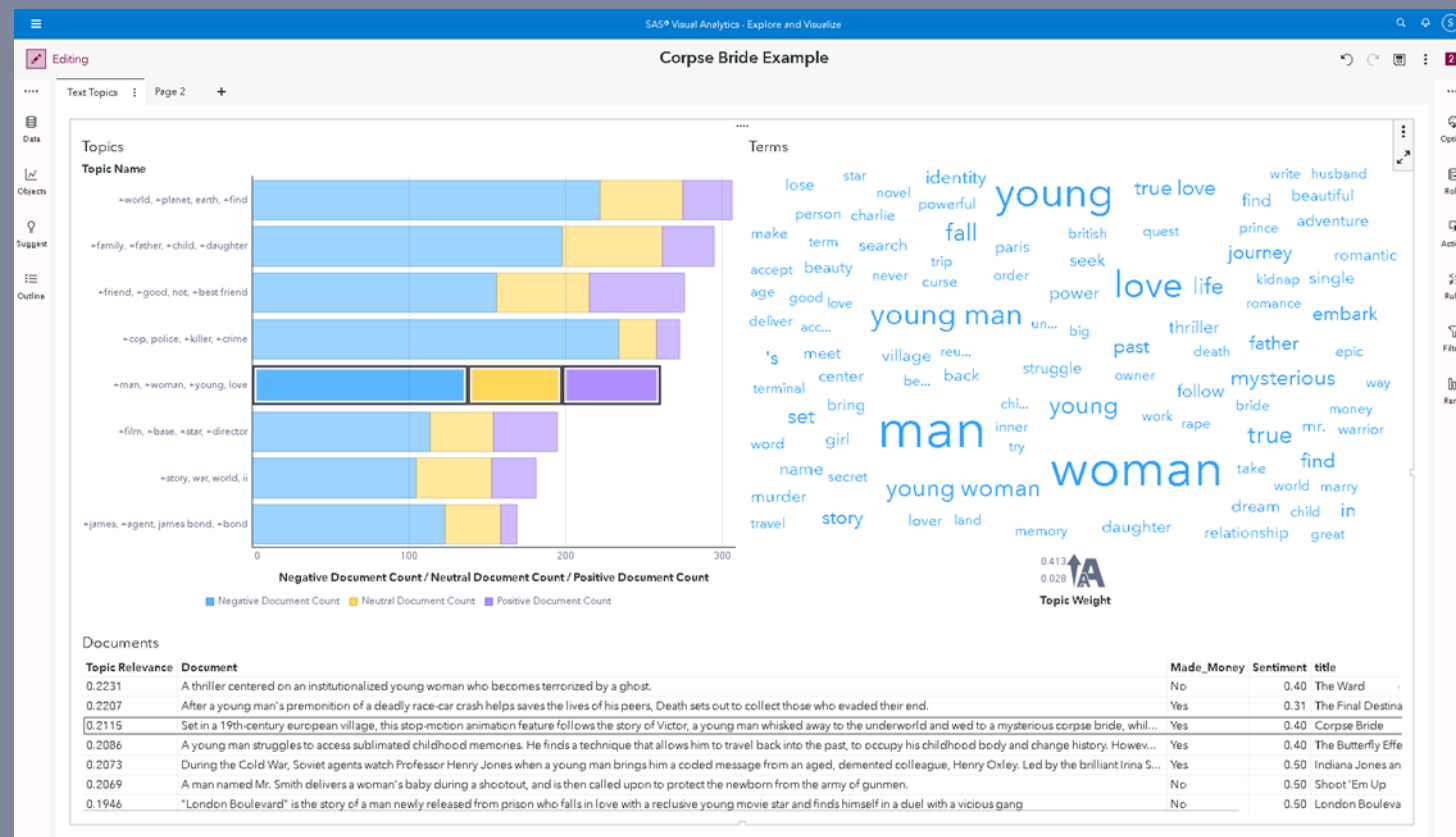
The + of Badges on the CV

[SAS EDUCATORS Portal](#)

Teacher Resource Portal

Totally free

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



Case - Movie Viewer Rating

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

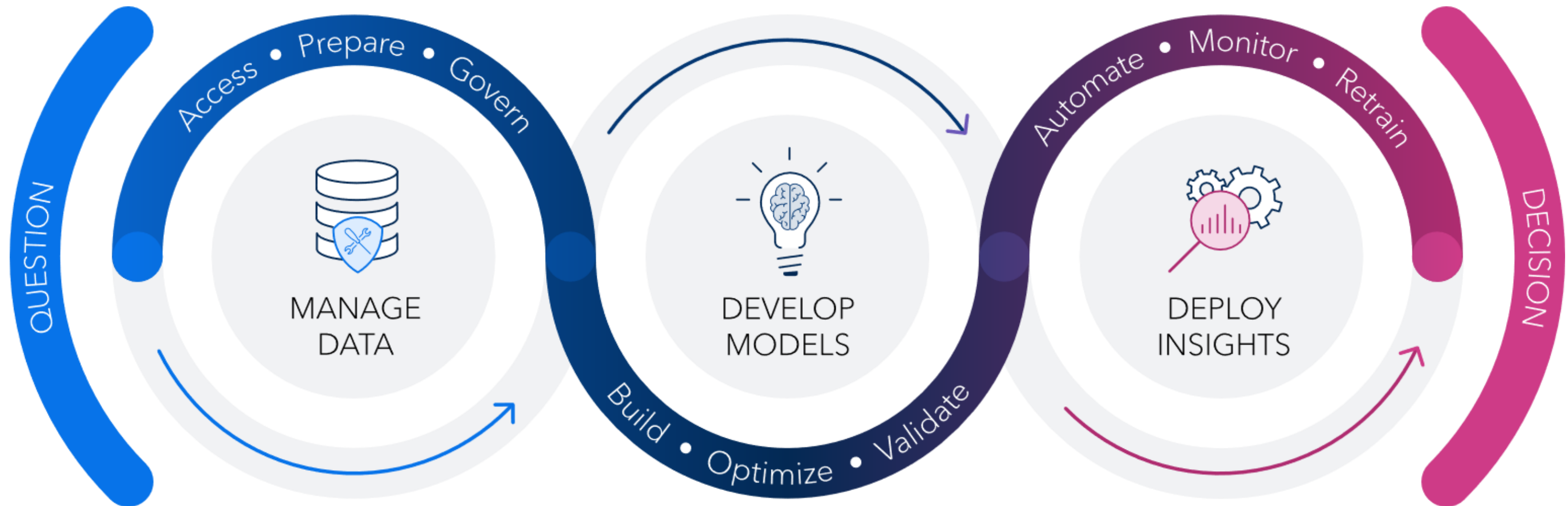
TF-IDF is used to determine how often a term appears in a large text or group of documents and therefore that term's importance to the document.

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.

Operationalizing Analytics

The Analytics Lifecycle

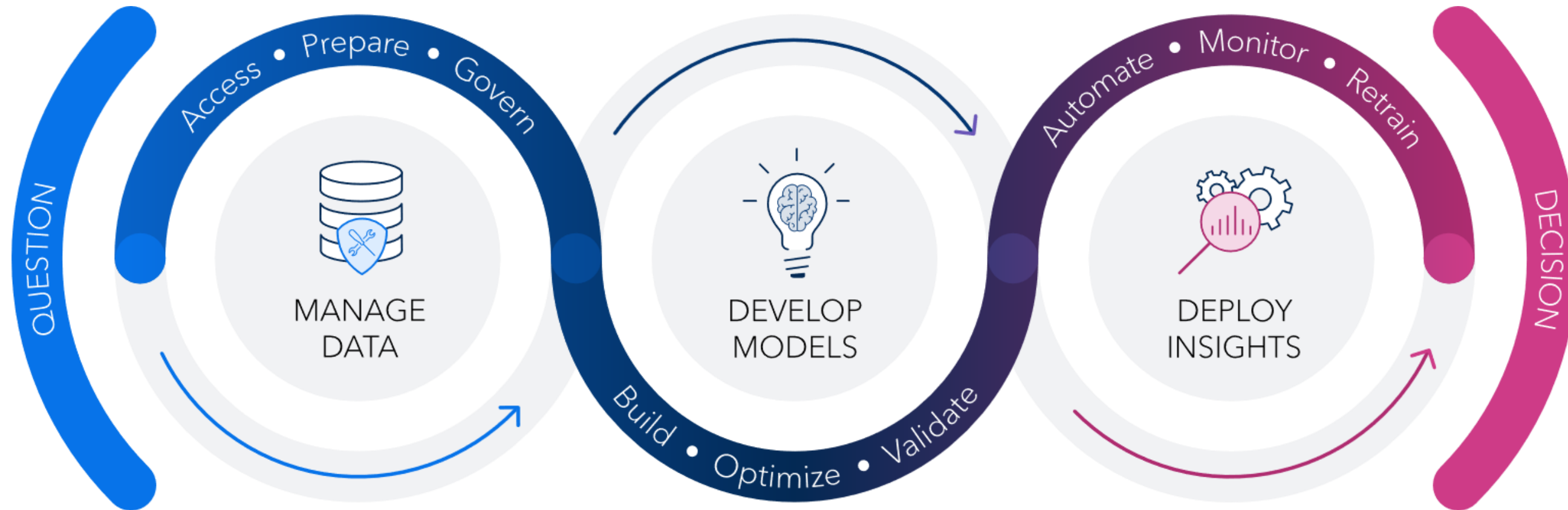


Sentiment Example: Customer review sentiment classifier

How can texted review data improve a predictive model?

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

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 preprocessing:
 - Tokenize, 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.



- Name: Movie Recommenders Unlimited, Inc.
- 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 analyzing 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

Data - Moviedata

Variables

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



What Everyone Should Know About The Movie Rating System.

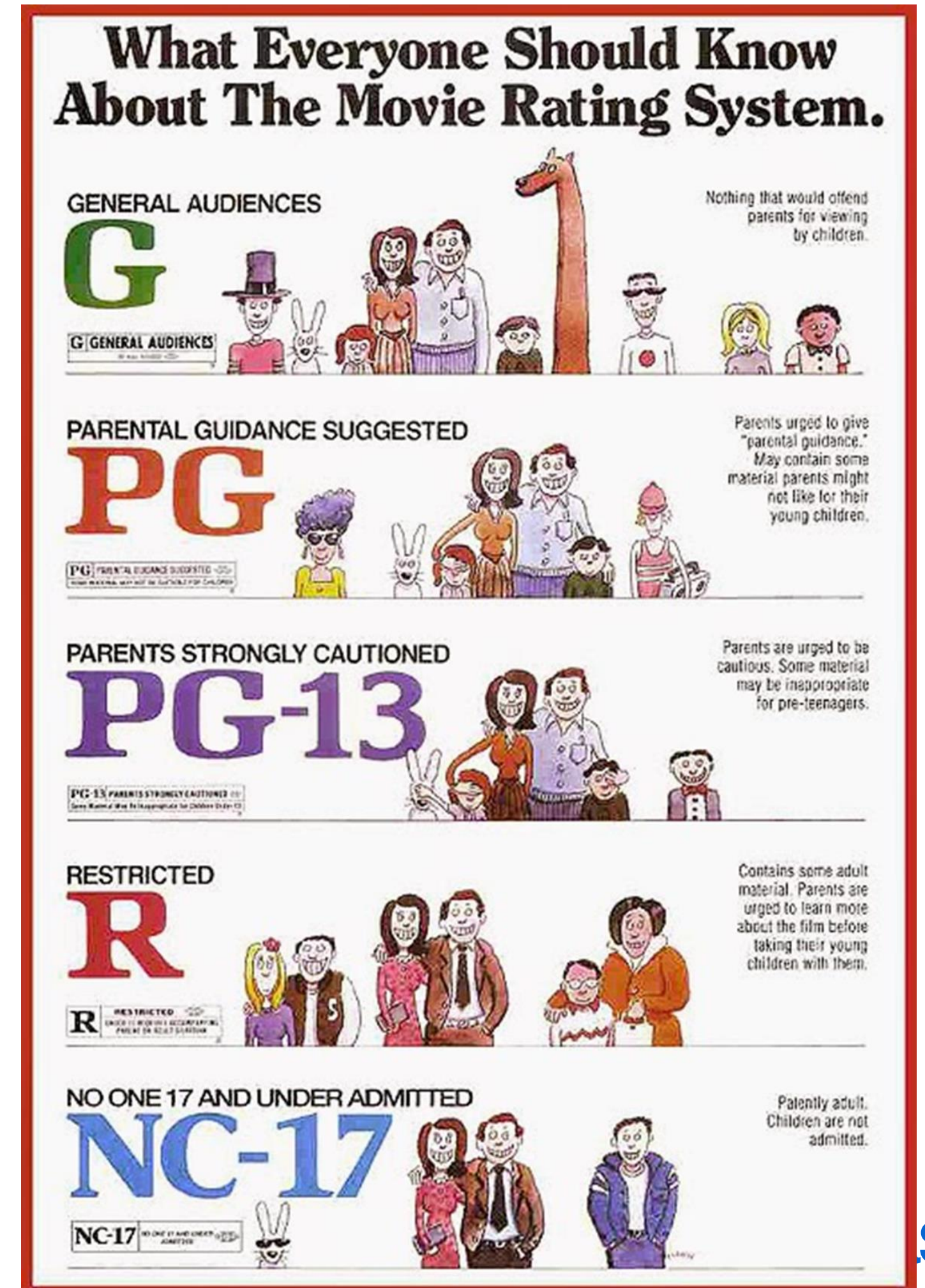
GENERAL AUDIENCES
G
Nothing that would offend parents for viewing by children.

PARENTAL GUIDANCE SUGGESTED
PG
Parents urged to give "parental guidance." May contain some material parents might not like for their young children.

PARENTS STRONGLY CAUTIONED
PG-13
Parents are urged to be cautious. Some material may be inappropriate for pre-teenagers.

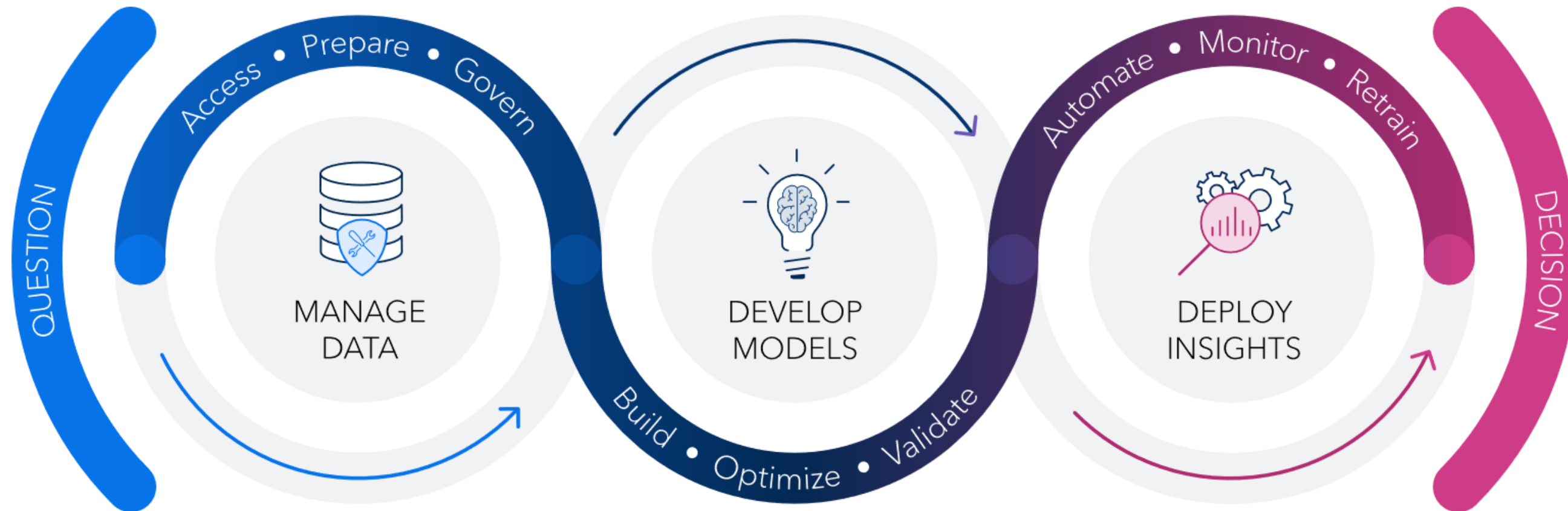
RESTRICTED
R
Contains some adult material. Parents are urged to learn more about the film before taking their young children with them.

NO ONE 17 AND UNDER ADMITTED
NC-17
Patently adult. Children are not admitted.



Operationalizing Analytics

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



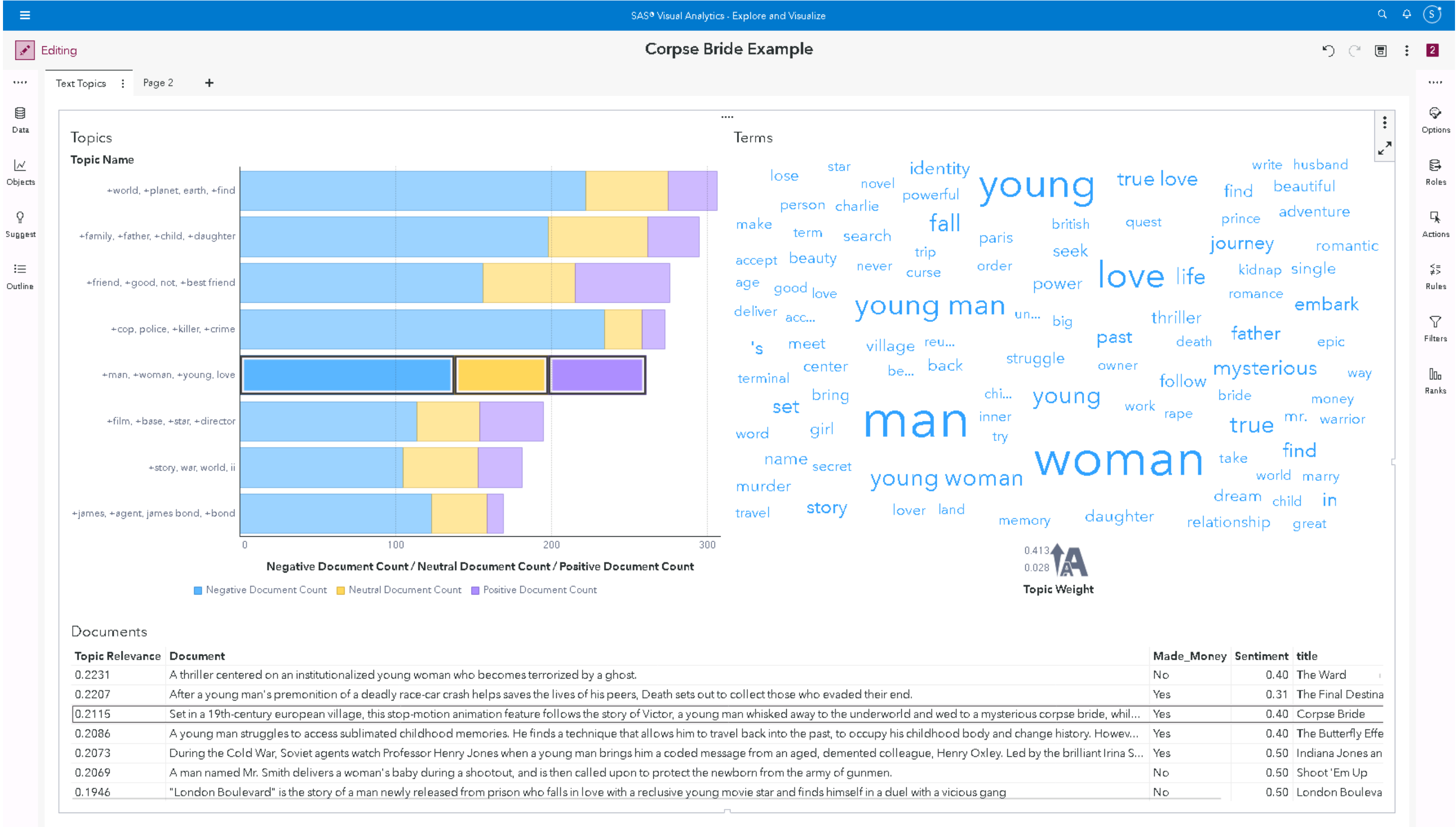
Decide which type of movie that maximizes the Viewer Rating

- 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 preprocessing:
 - 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...)

Meet the Text Topics Object



Data Roles

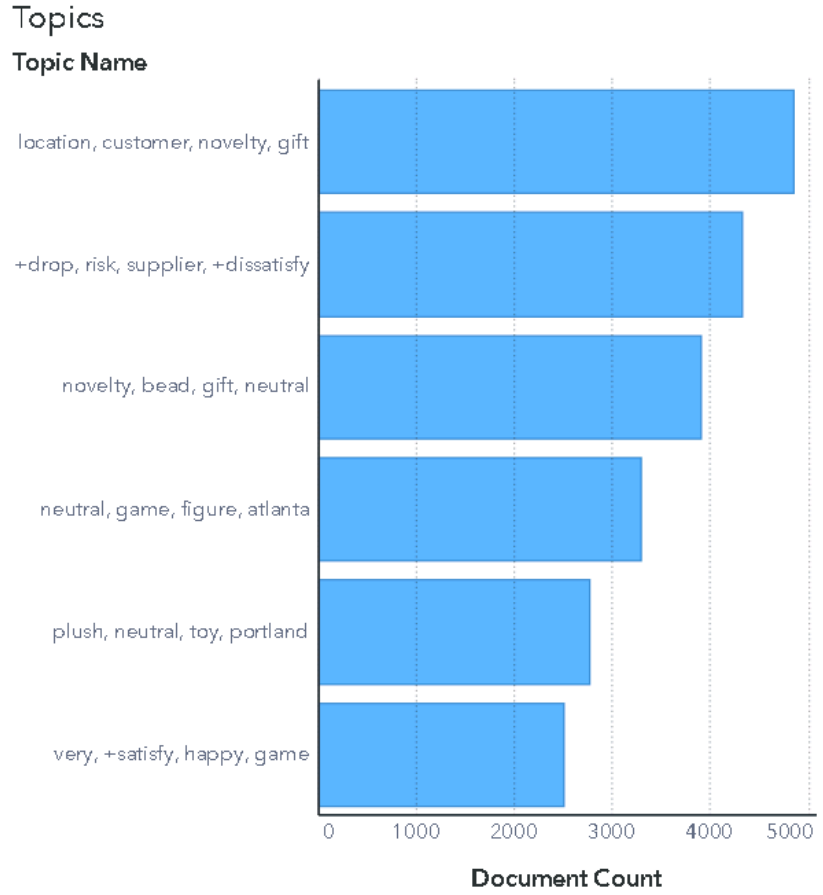
There must be a unique ID in your data!

The screenshot displays the SAS Text Topics interface. At the top, there are navigation elements: a back arrow, 'Text Topics', a vertical ellipsis, 'Text Topics Changed Options', 'Page 3', a right arrow, a plus sign, and a window icon. The main content area is divided into three panels:

- Topics:** A list of topic names with corresponding blue horizontal bars. The visible names are 'location, cu...', '+drop, risk,...', and 'novelty, be...'. There are four bars in total, with the fourth one being shorter.
- Terms:** A word cloud of terms. The most prominent terms are 'order', 'supplier', 'toy', and 'customer'. Other visible terms include 'basran', 'jacksonville', 'ok', 'portland', 'neutral', 'colorado springs', 'charlotte', 'hawaii', 'richmond', 'cleveland', 'raleigh', 'las vegas', 'madison', 'nashville', 'wichita', 'please', 'beach', 'hawaii', 'rock', 'minneapolis', 'miami', 'el paso', 'washington', 'st. louis', 'atlanta', 'mikeubee', 'st. louis', 'jefferson', 'new orleans', 'bimack', 'beach', 'plush', 'jackson', 'satisfy', 'columbus', 'des', 'concord', 'raleigh', 'promo', 'game', 'dissatisfy', 'dover', 'atlanta ready', 'charlotte', 'reno', 'beltsmore', 'place', 'locate', 'louisville', 'omaha', 'thrift', and 'hisco'.
- Data Roles:** A panel on the right with a title 'Data Roles' and a right arrow. It contains a dropdown menu set to 'Text topics - OrderNote 1'. Below it are two expandable sections: 'Document collection' containing 'OrderNote' with a document icon, and 'Document details' containing '+ Add'. At the bottom is another dropdown menu set to 'English'. On the far right of this panel is a vertical toolbar with icons for 'Options' (lightbulb), 'Roles' (database), 'Actions' (mouse cursor), 'Rules' (double arrows), and 'Filters' (funnel).

Output

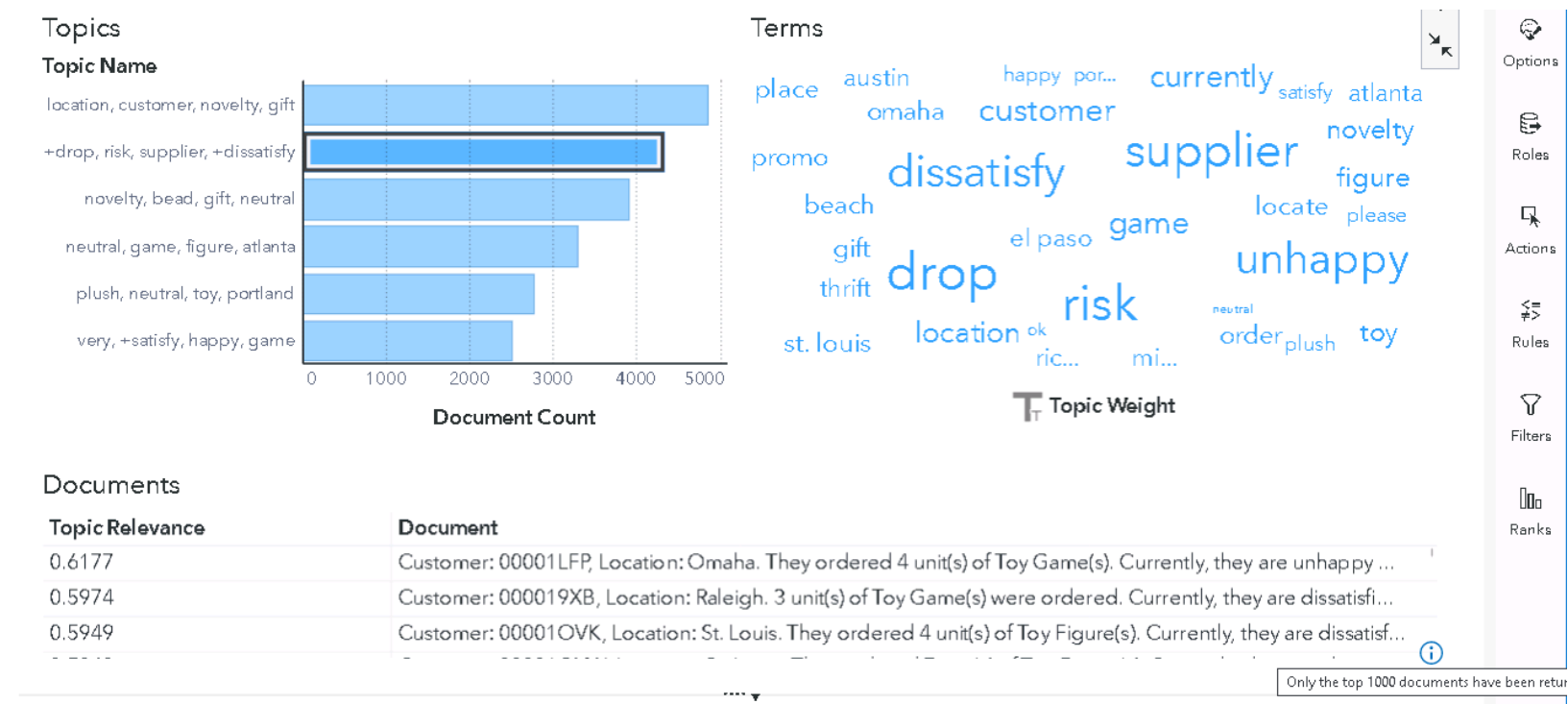
- Suggested topics
- Document count for each topic
- Word Cloud with word frequency



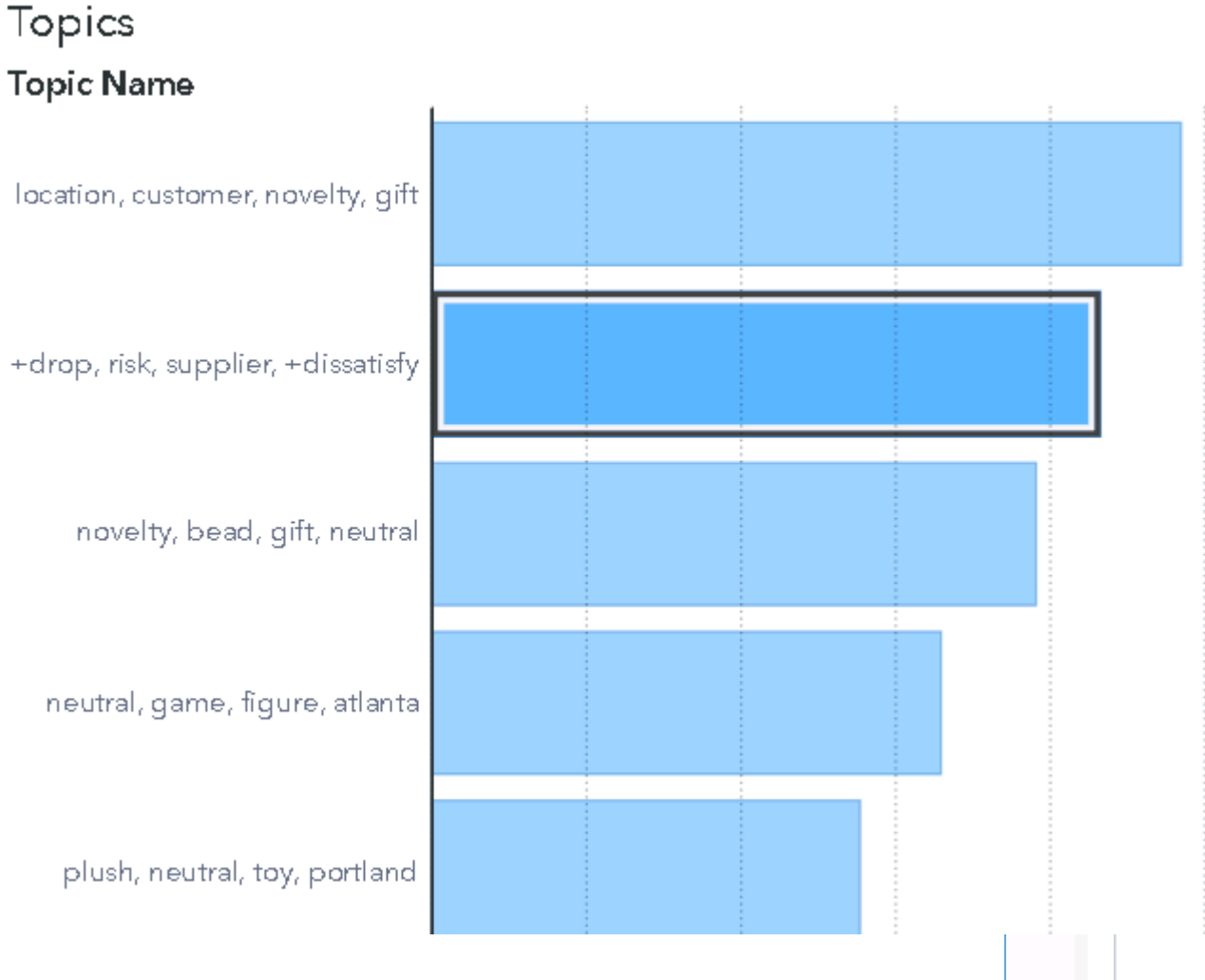
Selected Topic

Click on a topic

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



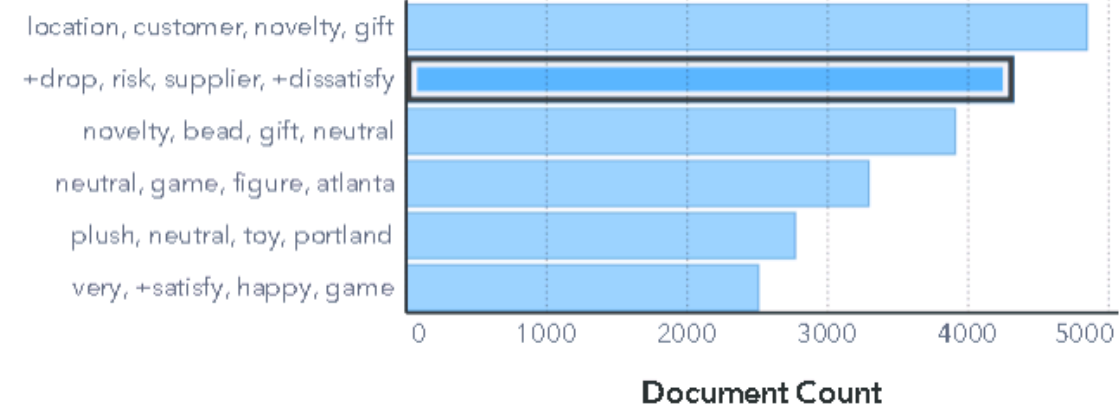
Term Weights can be positive and negative



Details Table

Topics

Topic Name



Terms



Documents

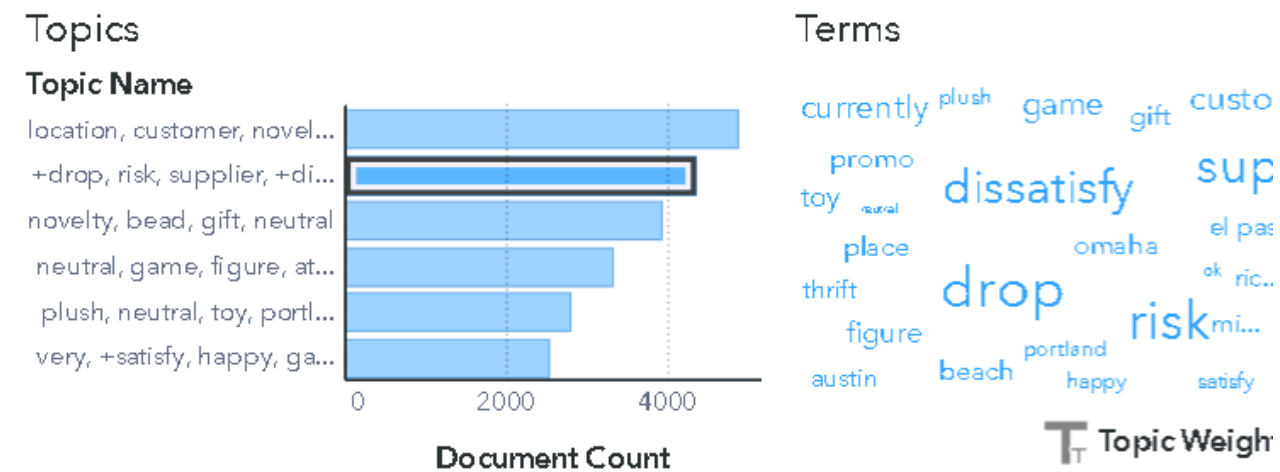
Topic Relevance	Document
0.6177	Customer: 00001LFP, Location: Omaha. They ordered 4 unit(s) of Toy Game(s). Currently, they are unhappy with us as ...
0.5974	Customer: 000019XB, Location: Raleigh. 3 unit(s) of Toy Game(s) were ordered. Currently, they are dissatisfied with u...
0.5949	Customer: 00001OVK,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

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)

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

Derive Topics

with a right mouse click

Topics

Topic Name

location, customer, novelty, gift
+drop, risk, supplier, +dissatisfy
novelty, bead, gift, neutral
neutral, game, figure, atlanta

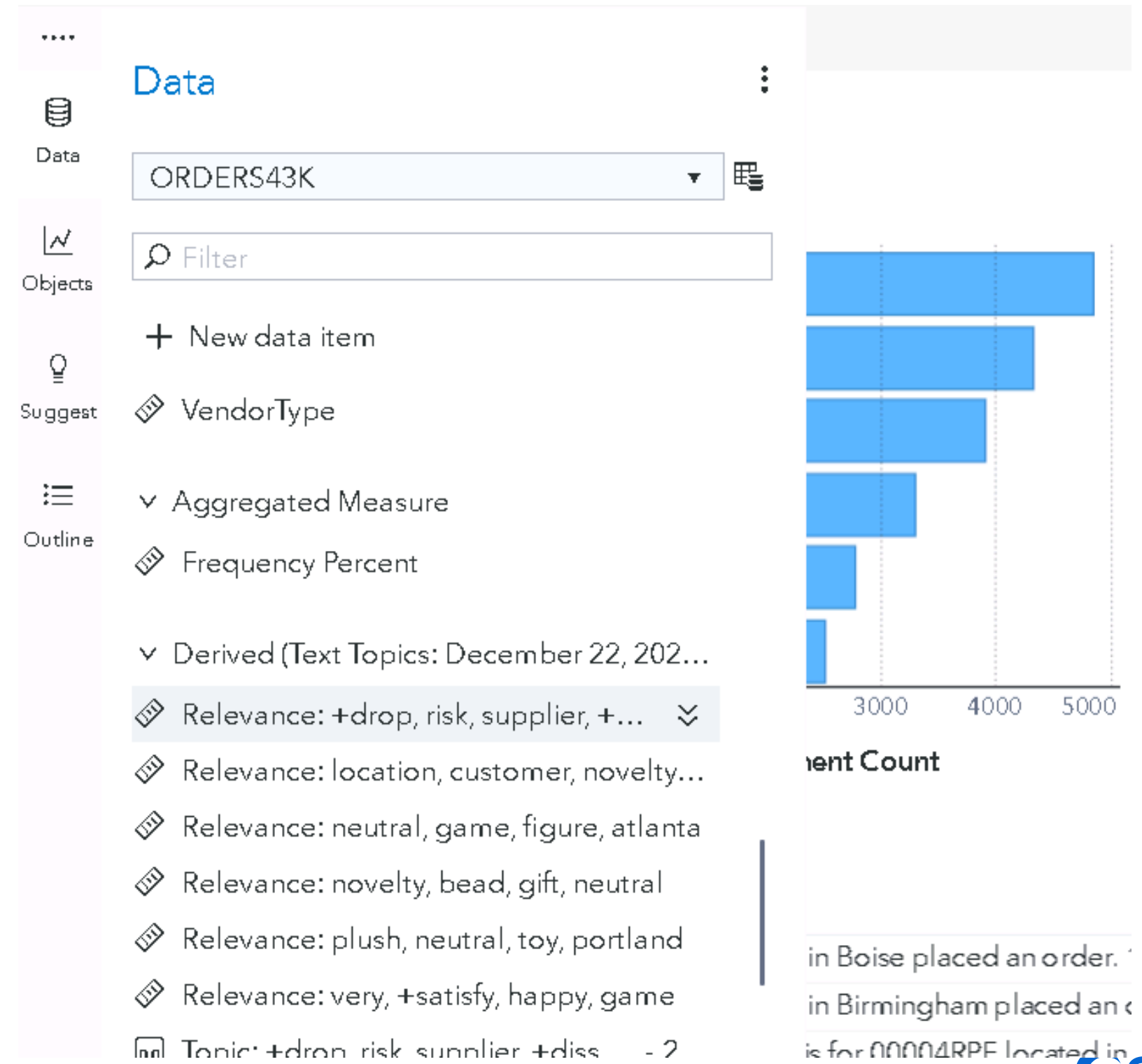
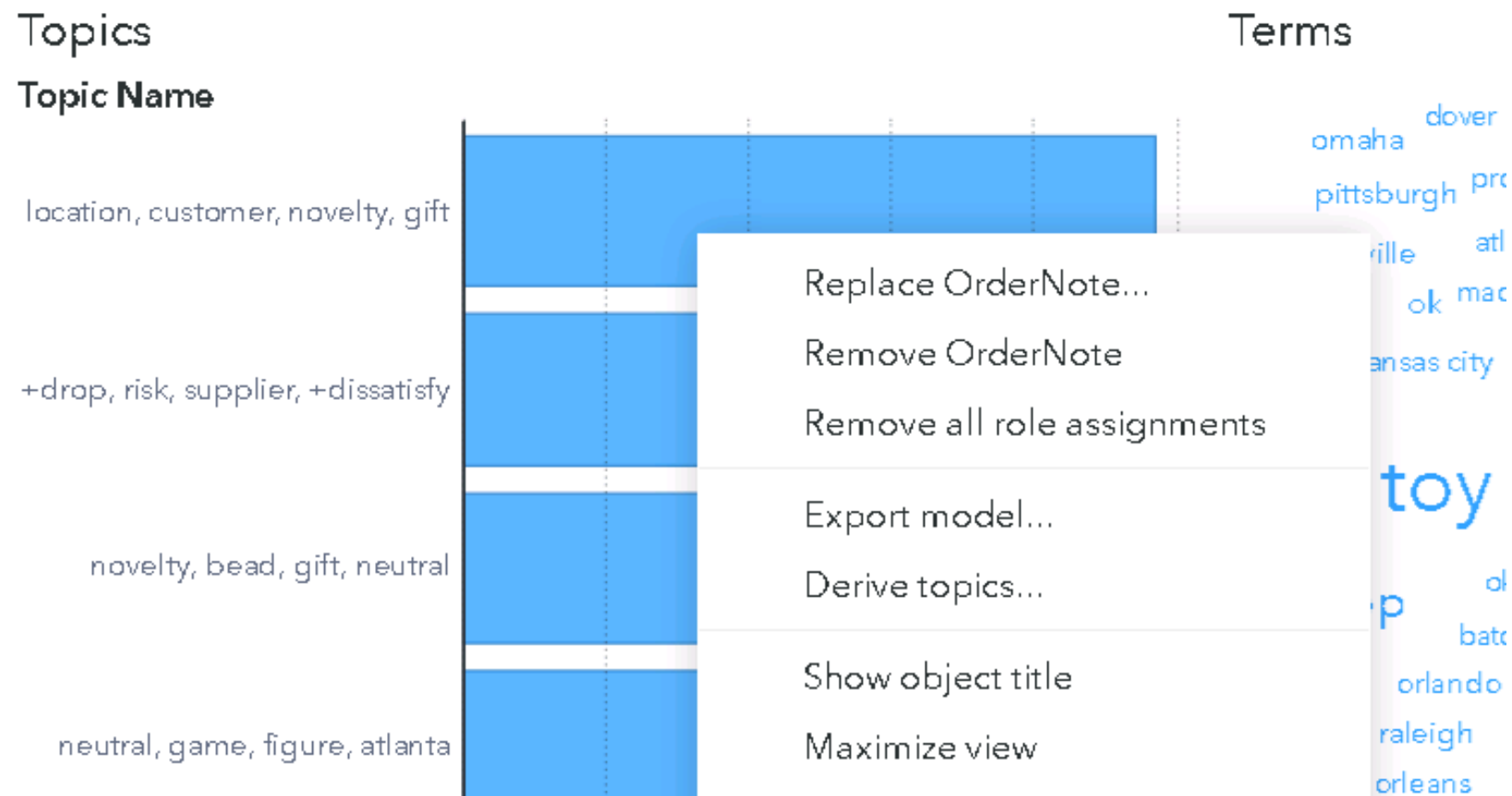
Terms

omaha dover
pittsburgh pre
ville atl
ok mac
ansas city
toy
p ol
batc
orlando
raleigh
orleans

- Replace OrderNote...
- Remove OrderNote
- Remove all role assignments
- Export model...
- Derive topics...
- Show object title
- Maximize view

Derive Topics

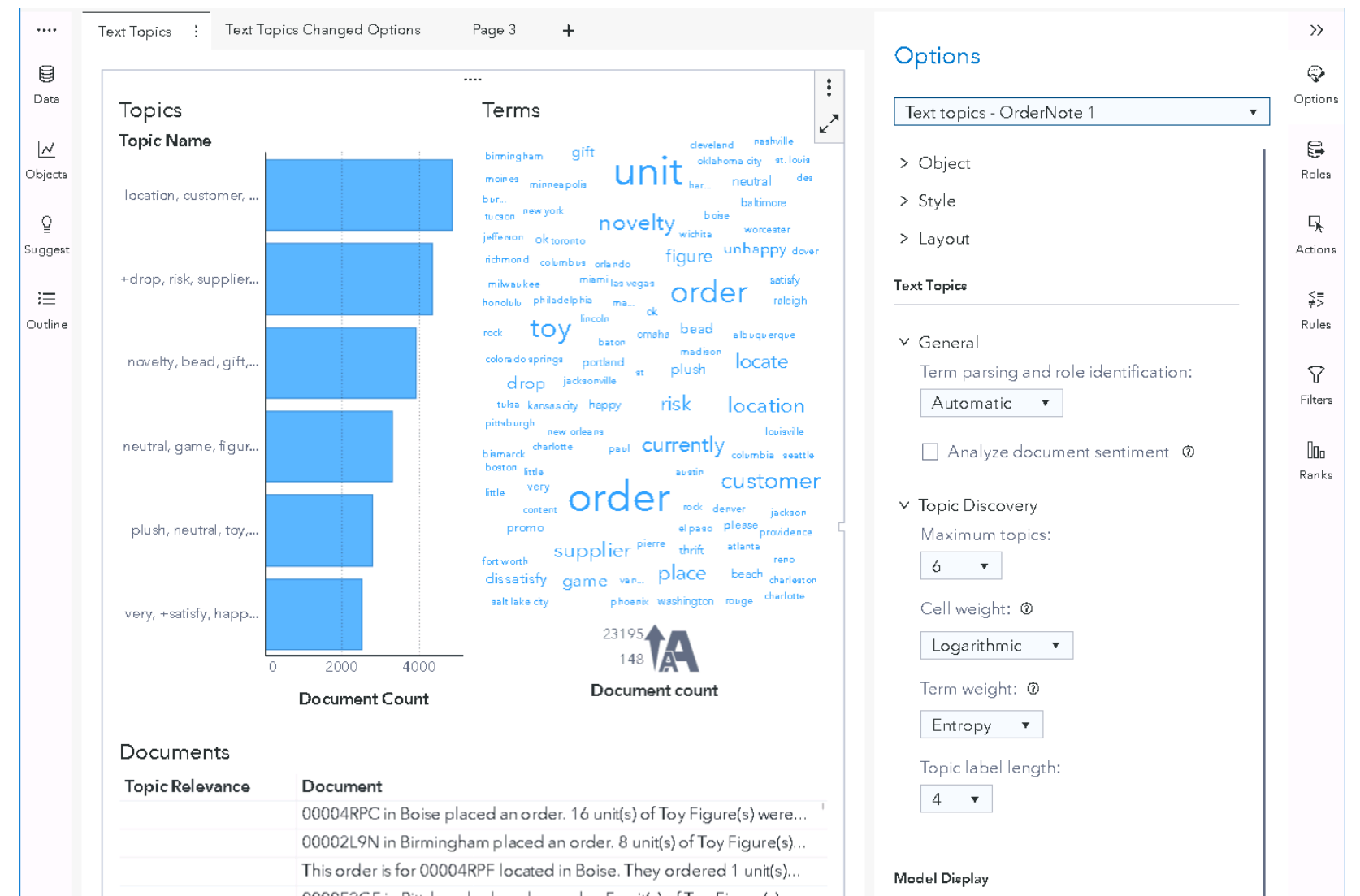
with a right mouse click



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



The screenshot displays the SAS Text Topics interface. The main window is titled "Text Topics Changed Options" and shows a "Topics" section with a horizontal bar chart and a "Terms" section with a word cloud. The "Topics" section lists six topics with their corresponding document counts. The "Terms" section shows a word cloud of terms associated with the topics. The "Options" panel on the right allows for adjusting various settings for the text topics.

Topic Name	Document Count
location, customer, ...	~4500
+drop, risk, supplier...	~3500
novelty, bead, gift,...	~3000
neutral, game, figur...	~2500
plush, neutral, toy,...	~2000
very, +satisfy, happ...	~1500

Options

Text topics - OrderNote 1

> Object

> Style

> Layout

Text Topics

General

Term parsing and role identification: Automatic

Analyze document sentiment

Topic Discovery

Maximum topics: 6

Cell weight: Logarithmic

Term weight: Entropy

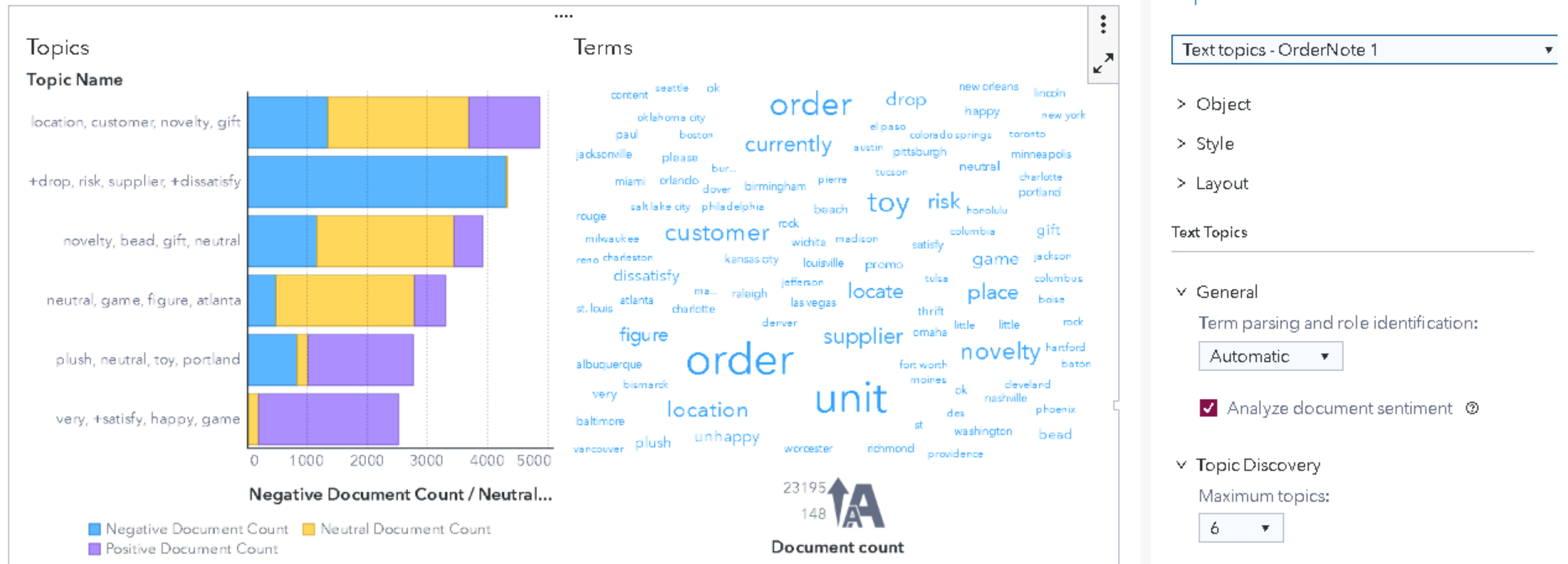
Topic label length: 4

Model Display

Documents

Topic Relevance	Document
	00004RPC in Boise placed an order. 16 unit(s) of Toy Figure(s) were...
	00002L9N in Birmingham placed an order. 8 unit(s) of Toy Figure(s)...
	This order is for 00004RPF located in Boise. They ordered 1 unit(s)...
	000052GE in Pittsburgh placed an order. 5 unit(s) of Toy Figure(s) w...

And if you do choose to analyze sentiments...



Spørsmål?

#learnsas

#skillbuilder

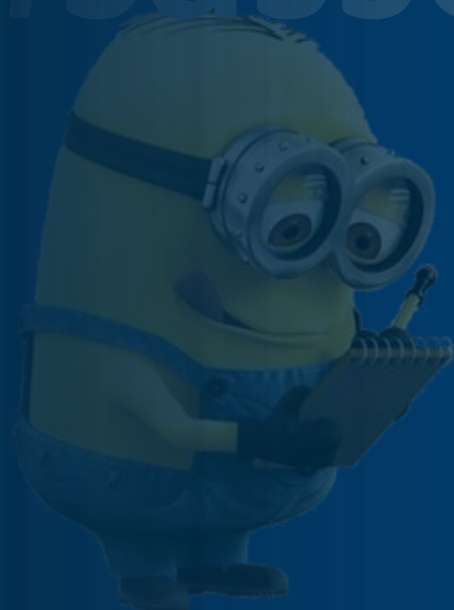
#lifelonglearner

#securethefuture



#sassoftware

#weareallacademics



Vegard Hansen

Academic Lead @ SAS

vegard.hansen@sas.com

www.linkedin.com/in/vegard-hansen/

