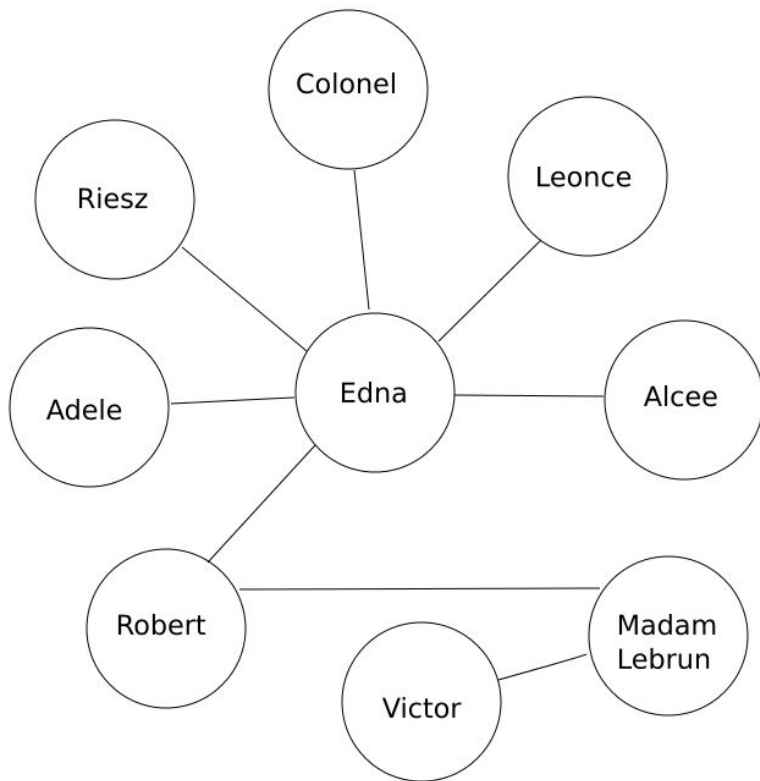


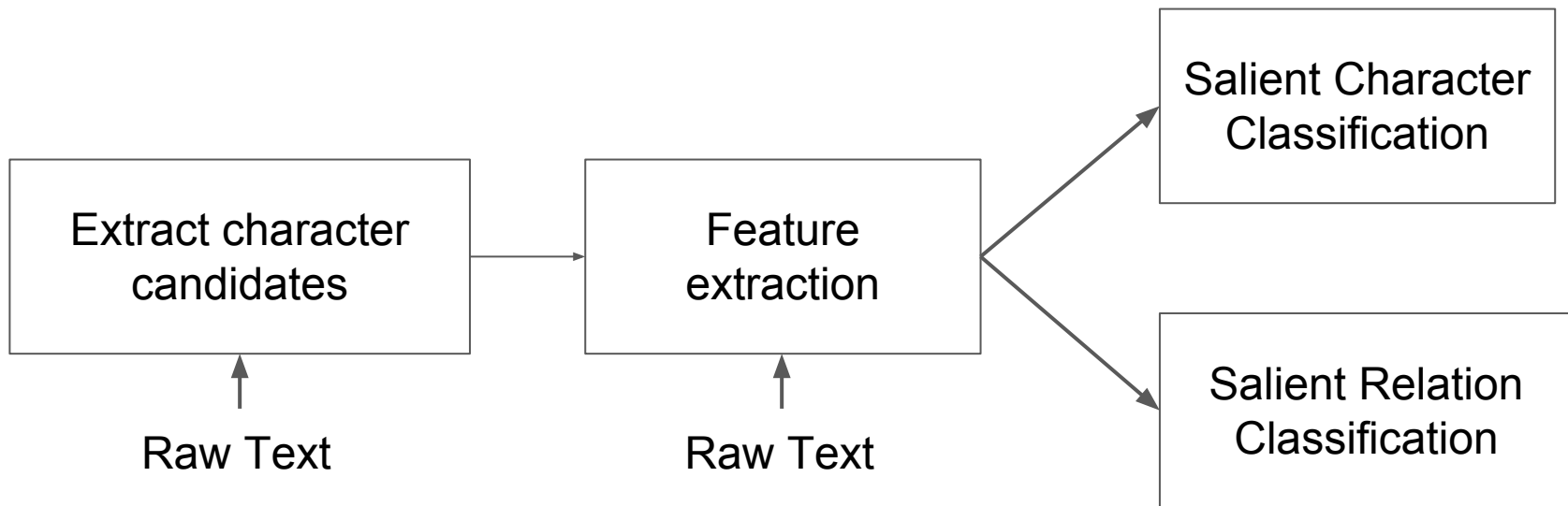
# Learning Character Graphs from Literature

Sumit Gogia, Min Zhang, Tommy Zhang

# Problem

Can we extract salient characters and whether they are significantly related from raw text?





# Data Collection

← Character List →

**Huckleberry Finn** - The protagonist and narrator of the novel. Huck is the thirteen-year-old son of the local drunk of St. Petersburg, Missouri, a town on the Mississippi River. Frequently forced to survive on his own wits and always a bit of an outcast, Huck is thoughtful, intelligent (though formally uneducated), and willing to come to his own conclusions about important matters, even if these conclusions contradict society's norms. Nevertheless, Huck is still a boy, and is influenced by others, particularly by his imaginative friend, Tom.

**Tom Sawyer** - Huck's friend, and the protagonist of *Tom Sawyer*, the novel to which *Huckleberry Finn* is ostensibly the sequel. In *Huckleberry Finn*, Tom serves as a foil to Huck: imaginative, dominating, and given to wild plans taken from the plots of adventure

## Characters

Huckleberry Finn

Tom Sawyer

...

## Relations

Huckleberry Finn

|

Tom Sawyer

...

Heuristically Extracted  
(person-like noun phrases)

### CANDIDATES

=====

```
('Well',)
('Tom',)
('the', 'king')
('the', 'duke')
('Huck',)
('Mary',)
('the', 'old', 'man')
('the', 'nigger')
('Buck',)
('the', 'widow')
('Tom', 'Sawyer')
...
```

# Feature Extraction

Tag Features: Capitalization and NER

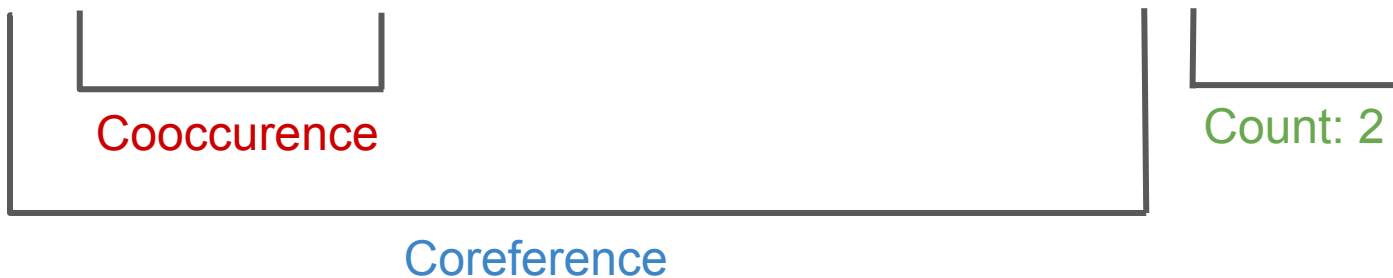
the widow **Bartley**

Coreference Features

Count Features

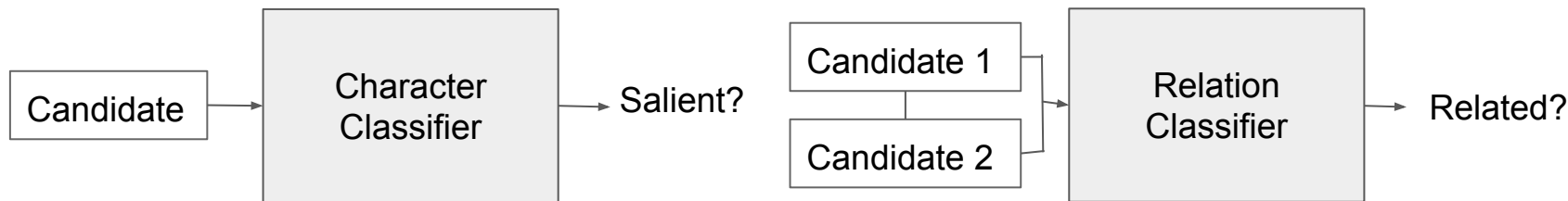
Cooccurrence Features

**Huckleberry** was **Tom**'s friend. But everyone called him **Huck**, **Huck** the incredible!



# Modeling and Inference

- Simple binary classification methods



- SVM, Random Forests
- Main adjustment is in large weights for positive data (lots of negative examples)

# Results (Character)

## Precision, Recall (Direct)

(‘Huck’, ‘Finn’) → 1  
(‘Huck’) → 1  
(‘Tom, Sawyer’) → 1  
(‘Aunt’, ‘Sally’) → 1  
(‘George’, ‘Jackson’) → 0

## Precision, Recall (Disambiguated)

(‘Huck’, ‘Finn’)  
(‘Huck’)

→ 1

(‘Tom, Sawyer’) → 1  
(‘Aunt’, ‘Sally’) → 1  
(‘George’, ‘Jackson’) → 0

	P1	R1	P2	R2
Baseline 1 (Top n-grams [20, 10, 5])	0.347	0.486	0.351	0.508
Baseline 2 (SVM, RBF, Count Features)	0.630	0.584	0.682	0.593
Full Feature Set (SVM, RBF)	0.638	0.702	0.684	0.593

??

# Results (Relations)

**Metrics:** Similar to those for characters!

	<b>P1</b>	<b>R1</b>
<b>Baseline 1 (Top co-occurring)</b>	0.303	0.198
<b>Baseline 2 (SVM, linear, cooc features)</b>	0.188	0.996
<b>Full Feature Set (SVM, linear)</b>	??	??

Very high bias (10)!

