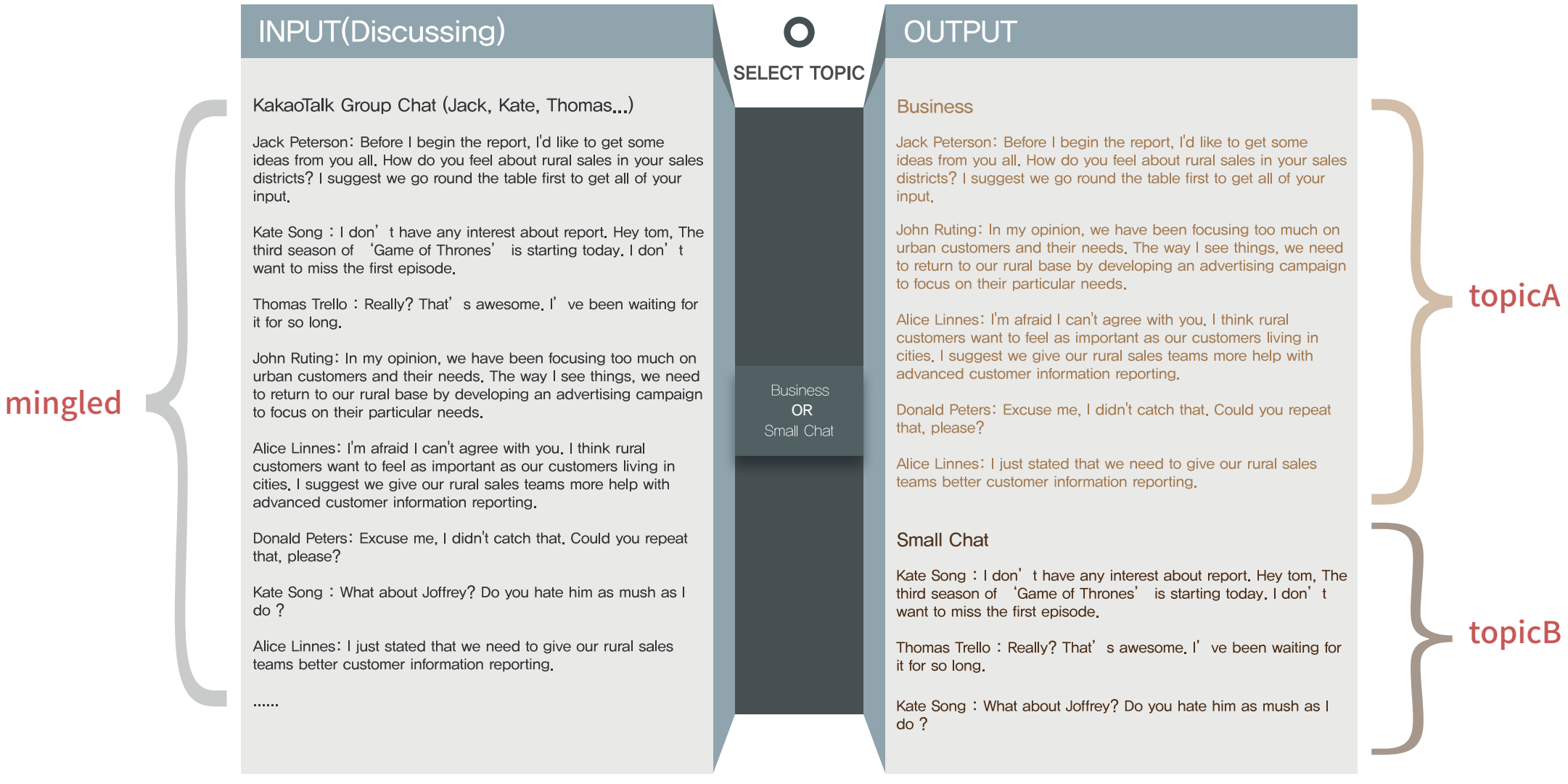
Natural Language Processing with Python Project 1 :

Categorizing distributed dialogues by interests

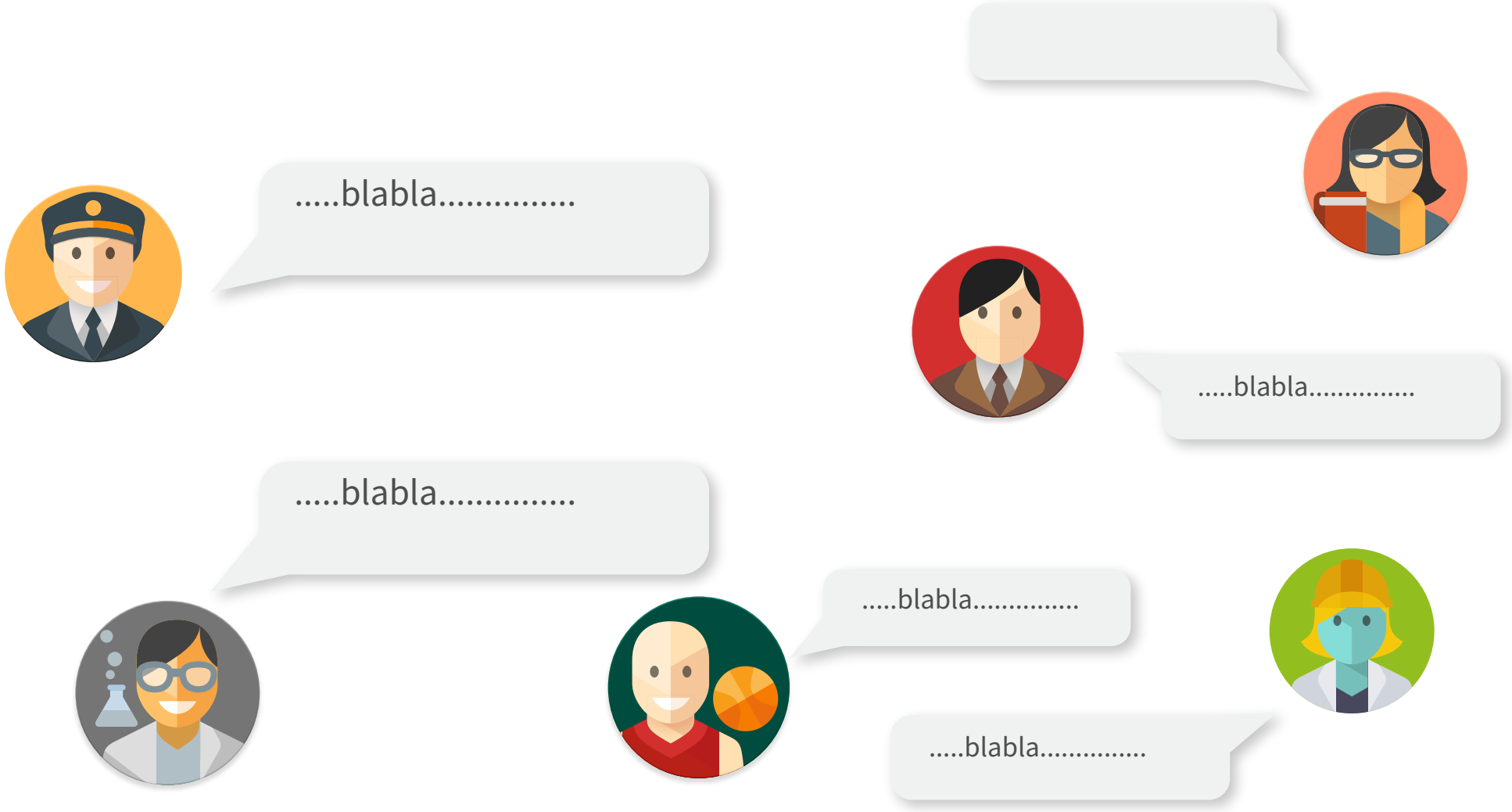
designed by

How can we deal with mingled-topic-dialogues?

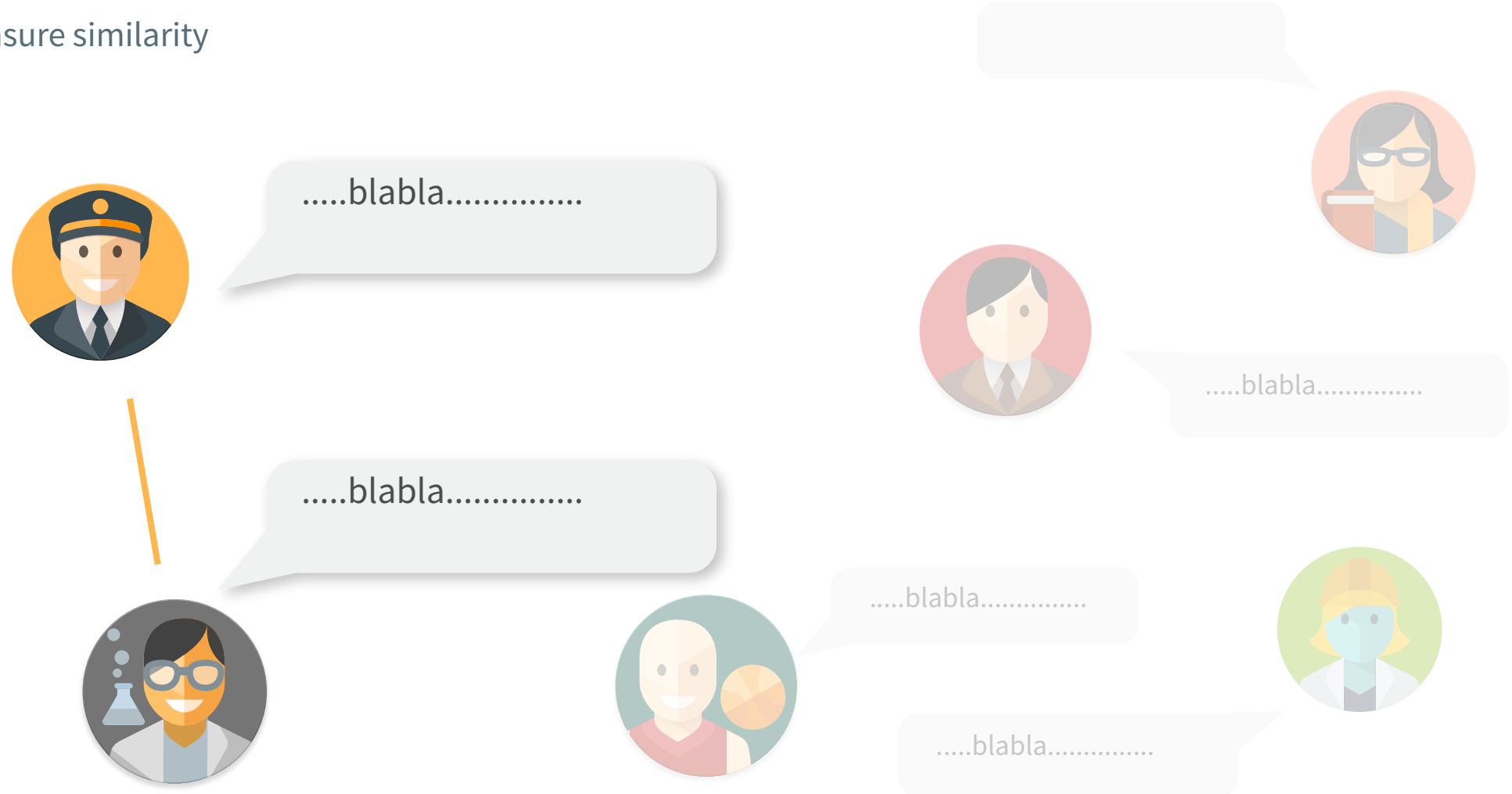

INPUT(Discussing)
SELECT TOPIC
OUTPUT
KakaoTalk Group Chat (Jack, Kate, Thomas...)
Jack Peterson: Before I begin the report, I'd like to get some ideas from you all. How do you feel about rural sales in your sales districts? I suggest we go round the table first to get all of your input.
Kate Song : I don't have any interest about report. Hey tom, The third season of 'Game of Thrones' is starting today. I don't want to miss the first episode.
Thomas Trello : Really? That's awesome. I've been waiting for it for so long.
John Ruting: In my opinion, we have been focusing too much on urban customers and their needs. The way I see things, we need to return to our rural base by developing an advertising campaign to focus on their particular needs.
Alice Linnes: I'm afraid I can't agree with you. I think rural customers want to feel as important as our customers living in cities. I suggest we give our rural sales teams more help with advanced customer information reporting.
Donald Peters: Excuse me, I didn't catch that. Could you repeat that, please?
Kate Song : What about Joffrey? Do you hate him as mush as I do ?
Alice Linnes: I just stated that we need to give our rural sales teams better customer information reporting.
.....
Business
OR
Small Chat
Business
Jack Peterson: Before I begin the report, I'd like to get some ideas from you all. How do you feel about rural sales in your sales districts? I suggest we go round the table first to get all of your input.
John Ruting: In my opinion, we have been focusing too much on urban customers and their needs. The way I see things, we need to return to our rural base by developing an advertising campaign to focus on their particular needs.
Alice Linnes: I'm afraid I can't agree with you. I think rural customers want to feel as important as our customers living in cities. I suggest we give our rural sales teams more help with advanced customer information reporting.
Donald Peters: Excuse me, I didn't catch that. Could you repeat that, please?
Alice Linnes: I just stated that we need to give our rural sales teams better customer information reporting.
Small Chat
Kate Song : I don't have any interest about report. Hey tom, The third season of 'Game of Thrones' is starting today. I don't want to miss the first episode.
Thomas Trello : Really? That's awesome. I've been waiting for it for so long.
Kate Song : What about Joffrey? Do you hate him as mush as I do ?
mingled
topicA
topicB

1. Link line to Characher

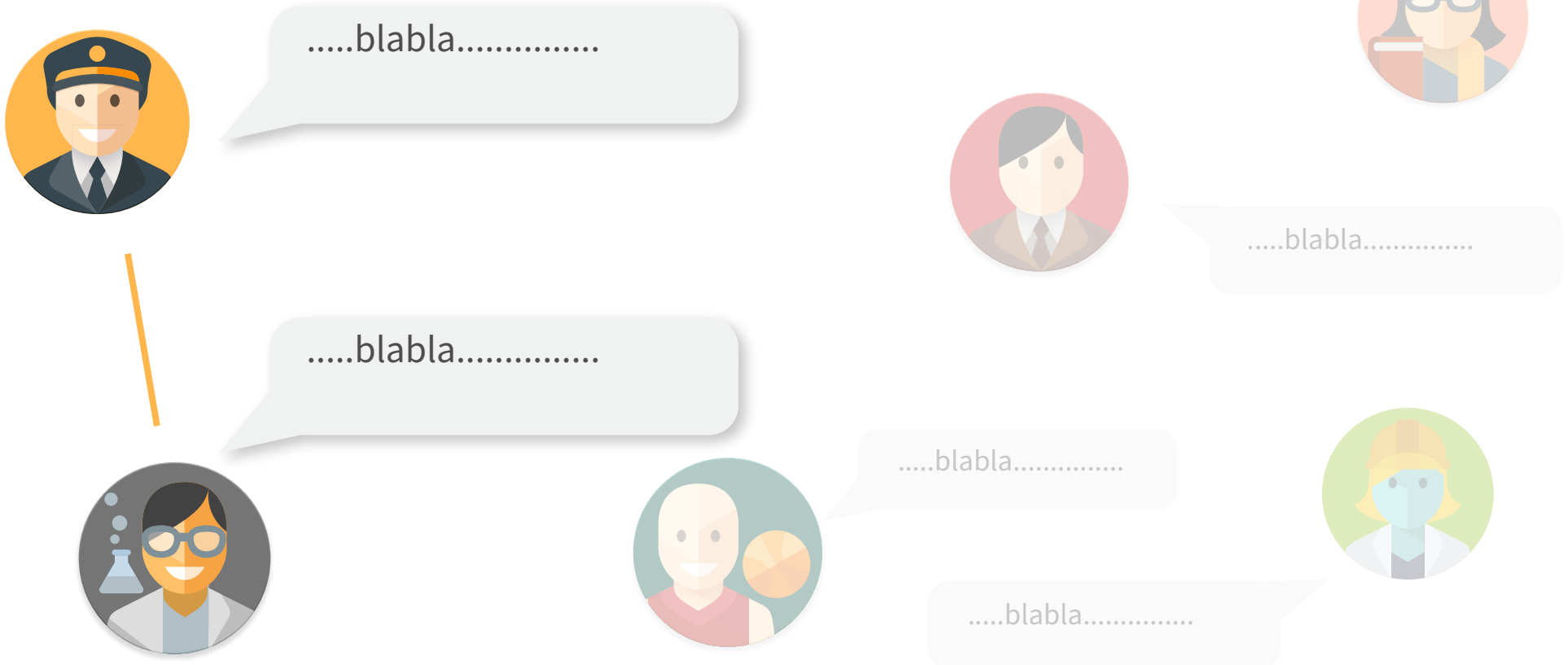
1. Link line to Characher

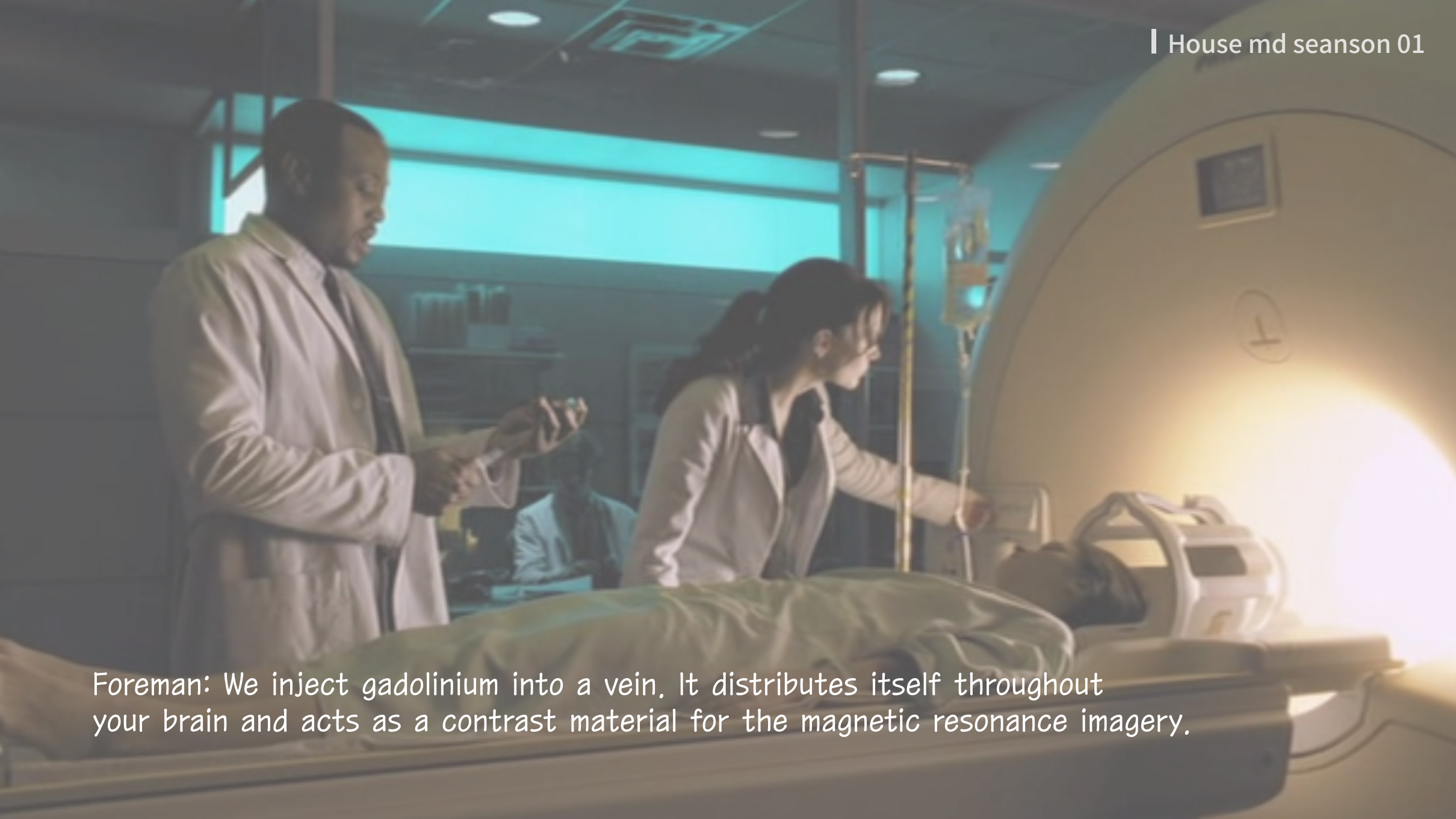


1. Link line to Characher
2. Measure similarity

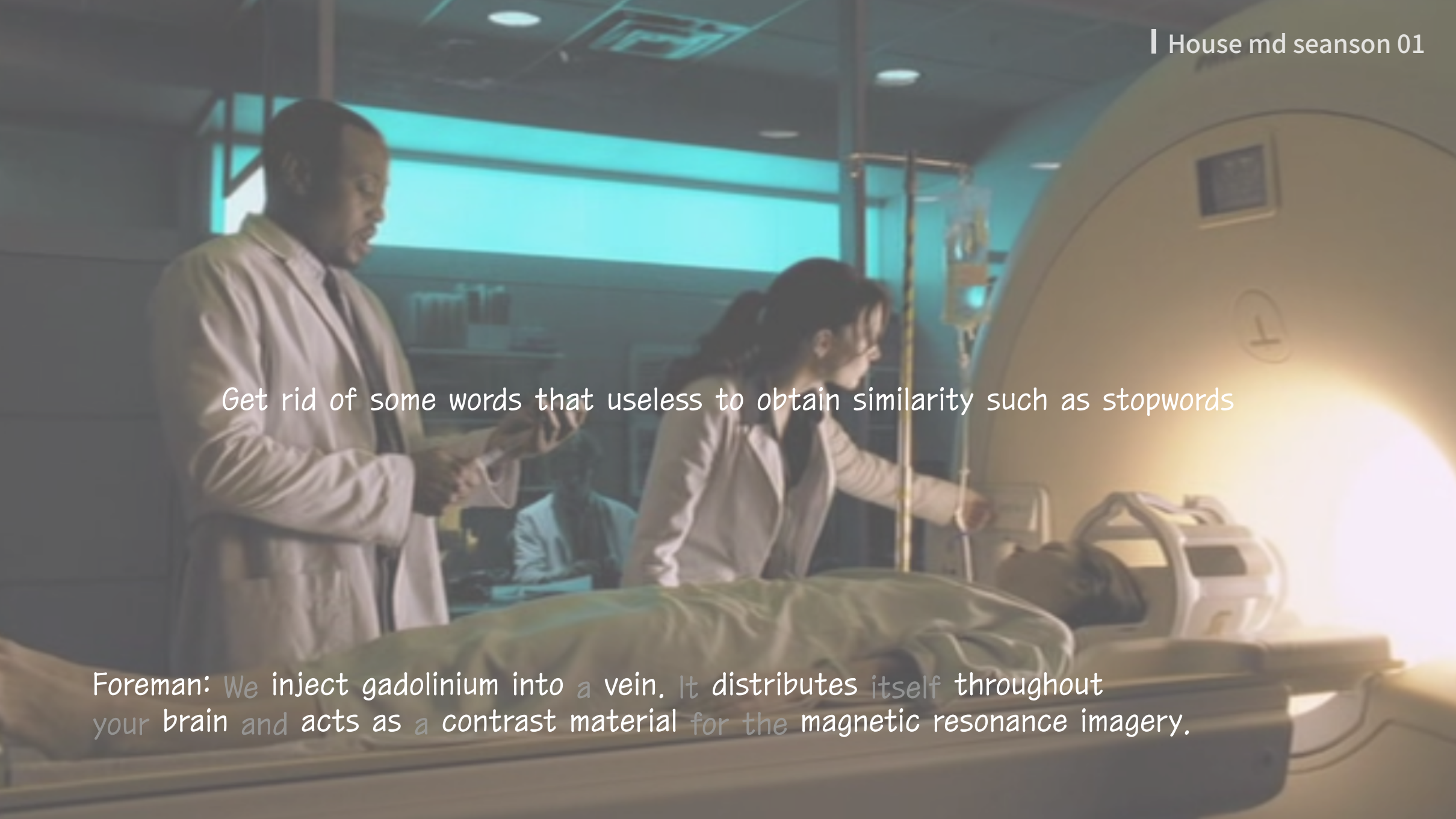


1. Link line to Characher
2. Measure similarity - modify charcter set
   - calculate same words between two



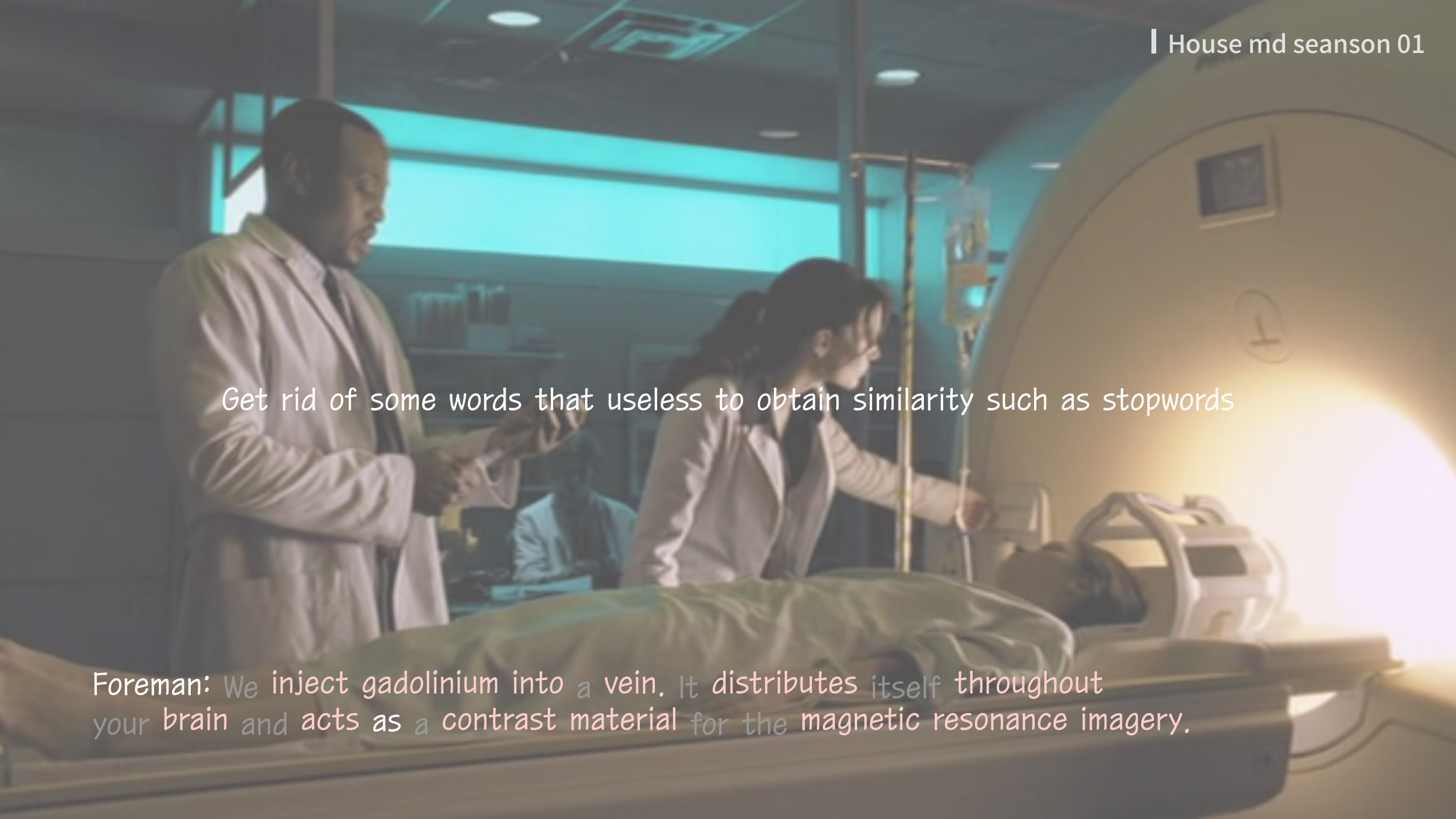


Foreman: We inject gadolinium into a vein. It distributes itself throughout your brain and acts as a contrast material for the magnetic resonance imagery.



Get rid of some words that useless to obtain similarity such as stopwords

Foreman: We inject gadolinium into a vein. It distributes itself throughout your brain and acts as a contrast material for the magnetic resonance imagery.



Get rid of some words that useless to obtain similarity such as stopwords

Foreman: We inject gadolinium into a vein. It distributes itself throughout your brain and acts as a contrast material for the magnetic resonance imagery.

1. Link line to Characher
2. Measure similarity - modify charcter set
   - calculate same words between two

%					
	100	70	30	13	13
	66	100	43	23	20
	35	40	100	18	5
	30	25	15	100	60
	18	15	7	54	100

1. Link line to Characher
2. Measure similarity
3. Divide into groups

%					
	100	70	30	13	13
	66	100	43	23	20
	35	40	100	18	5
	30	25	15	100	60
	B	15	7	54	100

A

1. Link line to Characher
2. Measure similarity
3. Divide into groups

%					
	100	70	30	13	13
	66	100	43	23	20
	35	40	100	18	5
	30	25	15	100	60
	B	B	7	54	100

A

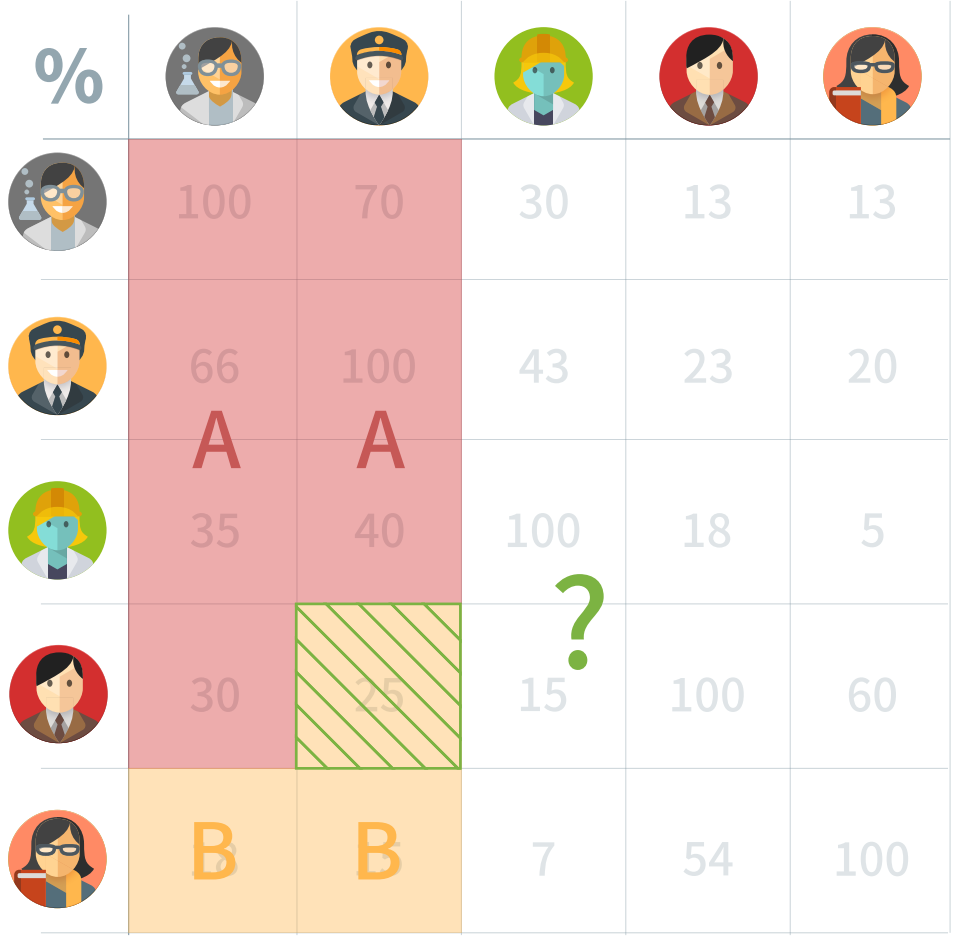
A

B

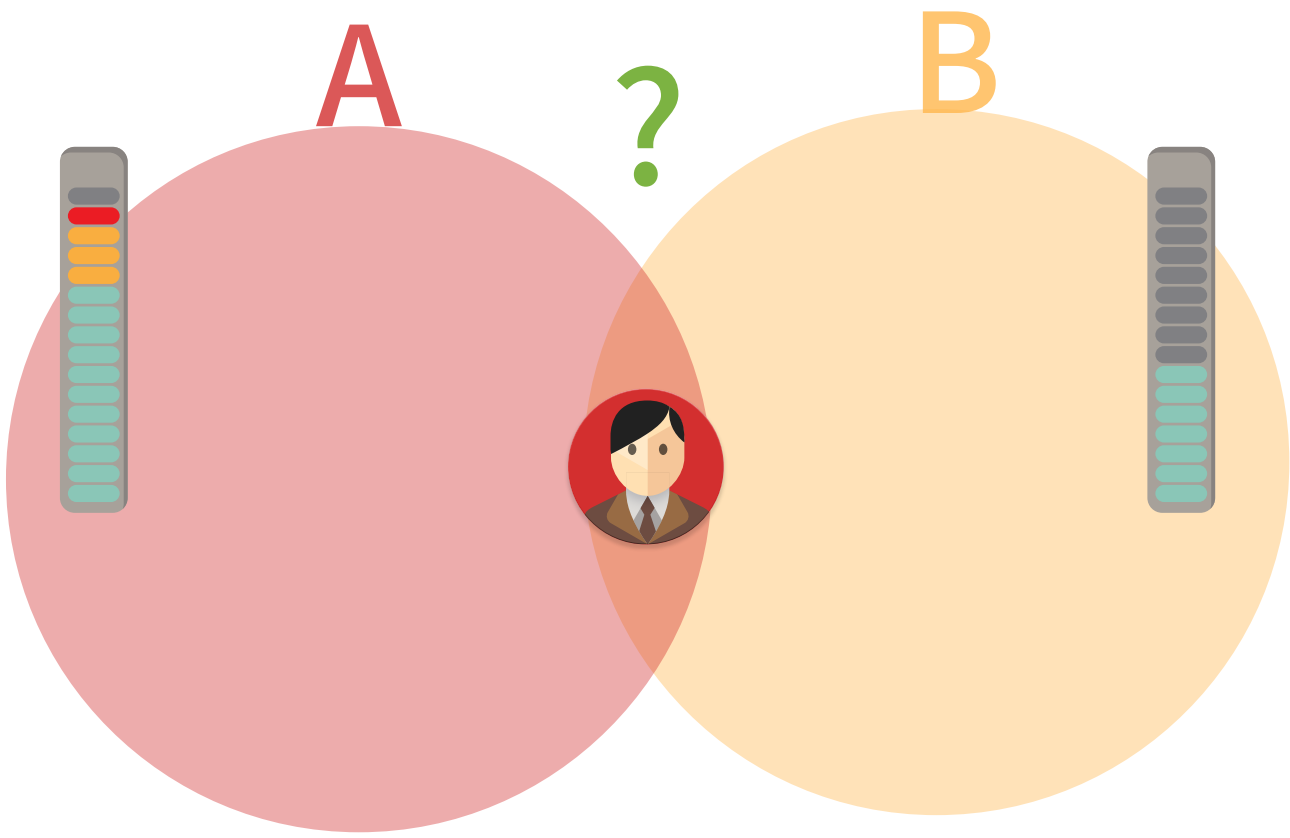
1. Link line to Characher
2. Measure similarity
3. Divide into groups

%					
	100	70	30	13	13
	66	100	43	23	20
	A	A			
	35	40	100	18	5
	30	25	15	100	60
	B	B	7	54	100

1. Link line to Characher
2. Measure similarity
3. Divide into groups



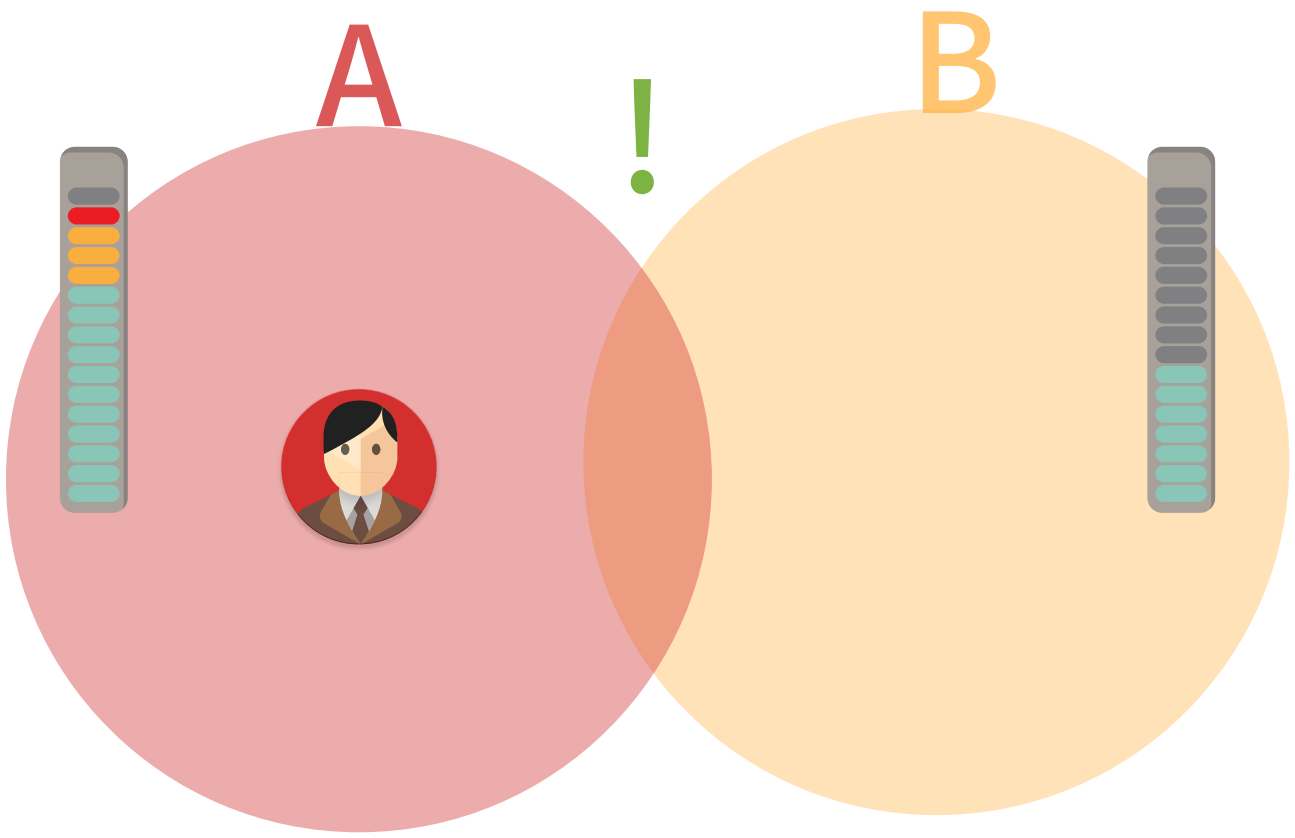
1. Link line to Characher
2. Measure similarity
3. Divide into groups - Processing ambigious set



avarage path-similarity of word that he/she said

from Wordnet

1. Link line to Characher
2. Measure similarity
3. Divide into groups - Processing ambigious set



avarage path-similarity of word that he/she said

from Wordnet

1. Link line to Characher
2. Measure similarity
3. Divide into groups - Processing ambigious set

setA setB

firstDividedSet = [['MONICA', 'RACHEL', 'ROSS', 'JOEY', 'CHANDLER'], ['Cuddy', 'Cameron', 'Rebecca', 'Wilson', 'Chase']]

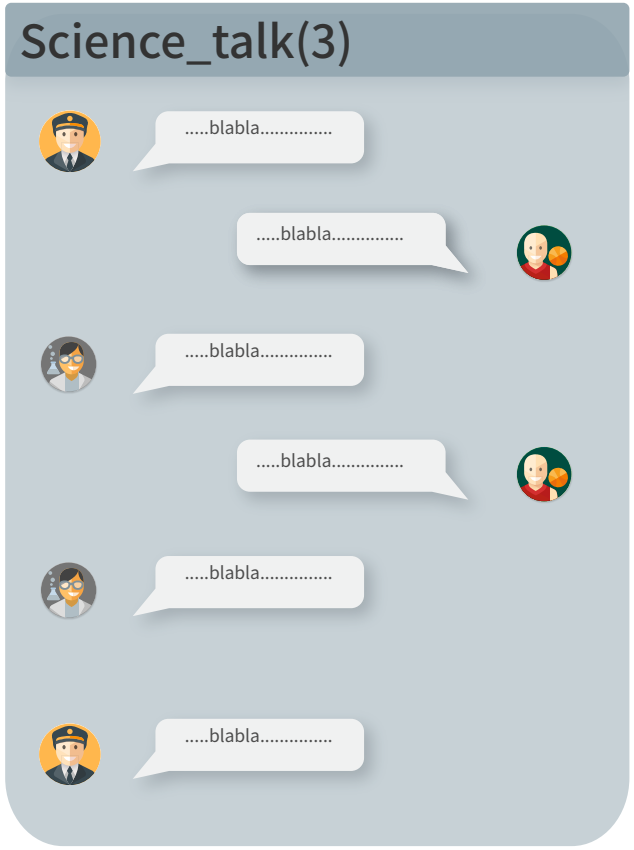
ambigSet = ['PHOEBE', 'PAUL', 'House', 'Foreman']

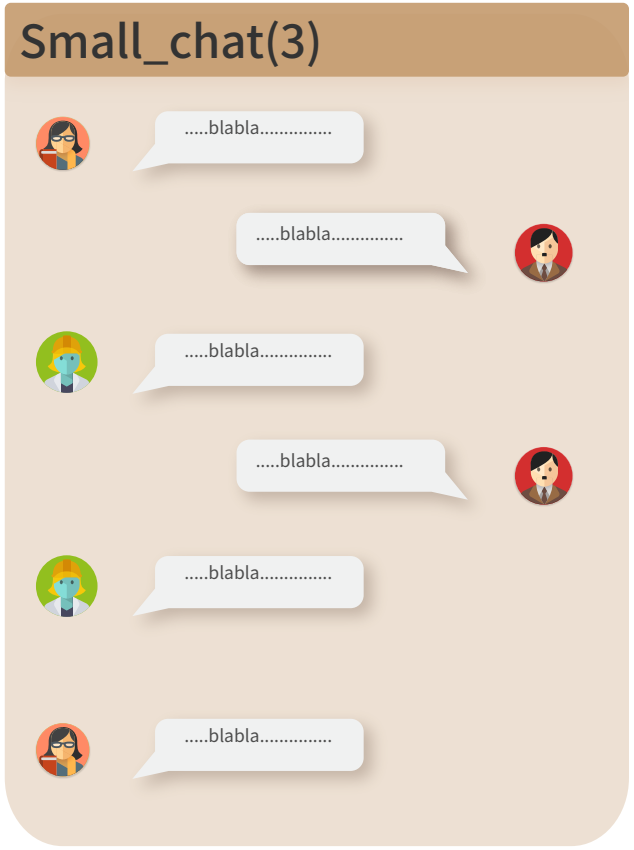
print chsSetByScr(setA, setB, ambigSet, scoreDic)

PHOEBE	0.073	0.075
PAUL	0.087	0.086
House	0.078	0.083
Foreman	0.078	0.083

['MONICA', 'RACHEL', 'ROSS', 'JOEY', 'CHANDLER', 'PAUL'], ['Cuddy', 'Cameron', 'Rebecca', 'Wilson', 'Chase', 'PHOEBE', 'House', 'Foreman']

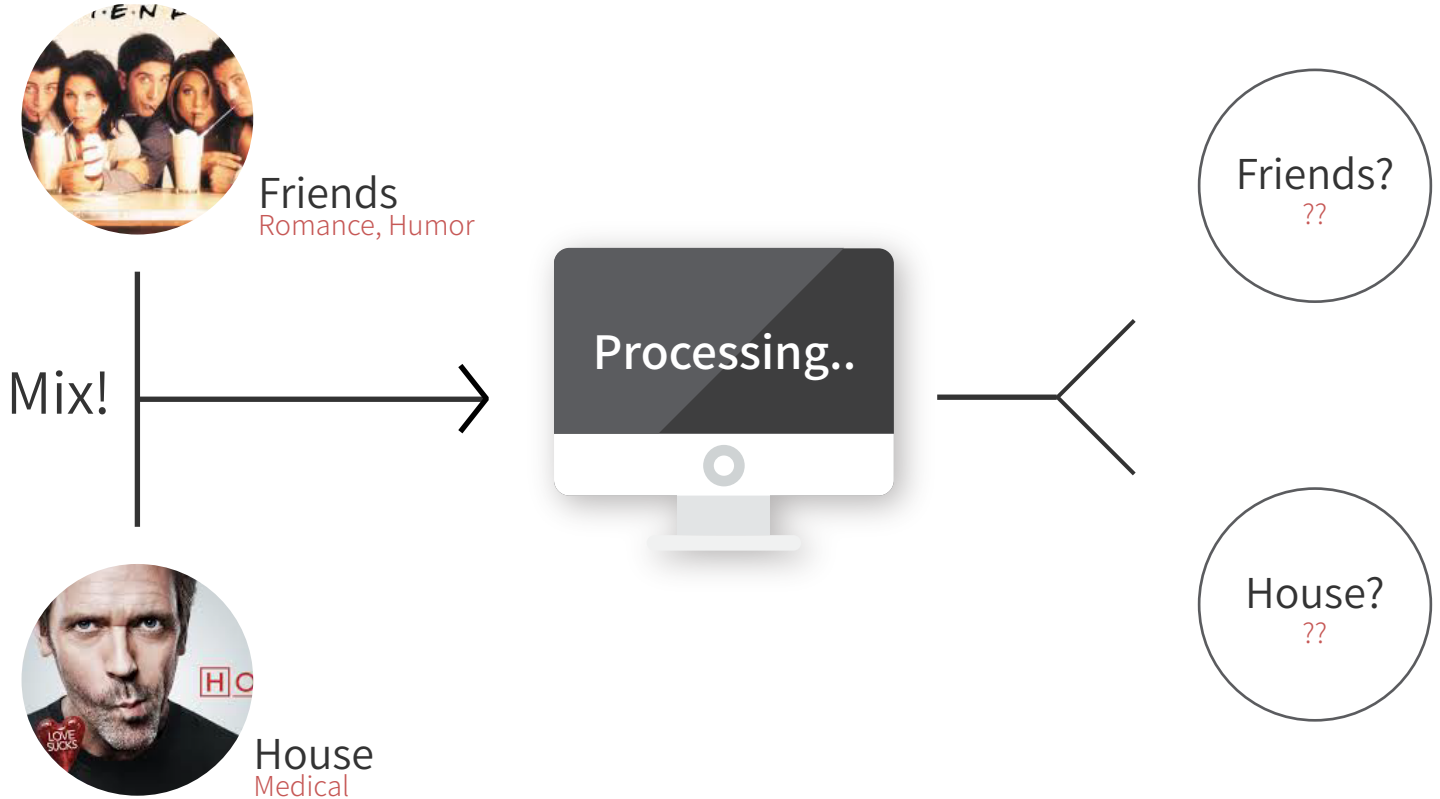
1. Link line to Characher
2. Measure similarity
3. Divide into groups
4. Get result





Mixed dialogues into one text as an input.

Seperated them as a result in our algorithm

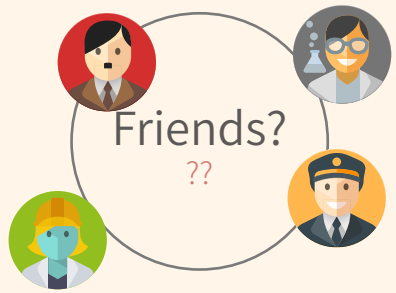


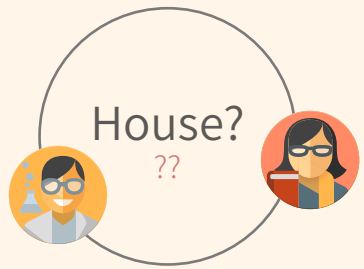
To measure quality of our program, we built the function that check whether the person is in the right place

Friends
Romance, Humor

House
Medical

VS.



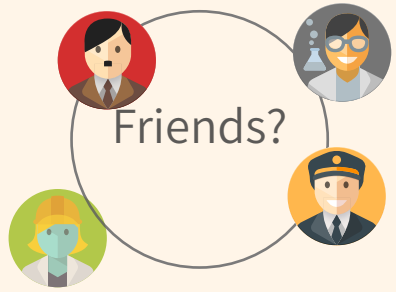


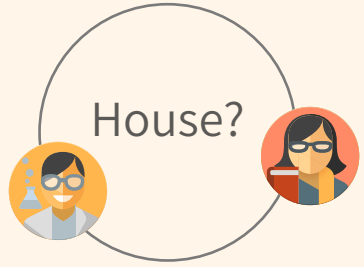
$$\frac{\text{right people}}{\text{total people}} \times 100$$

Answer

VS.

Result





Test Input : Scripit of House md season 1_01 and Friends season1_01

Our result

[['MONICA', 'RACHEL', 'ROSS', 'JOEY', 'CHANDLER', 'PAUL'],
['Cuddy', 'Cameron', 'Rebecca', 'Wilson', 'Chase', 'PHOEBE', 'House', 'Foreman']]

92.8%

Golden set

[['MONICA', 'RACHEL', 'PHOEBE', 'PAUL', 'JOEY', 'CHANDLER', 'ROSS'],
['Cuddy', 'Wilson', 'Foreman', 'House', 'Cameron', 'Rebecca', 'Chase']]

Thank you :)