This notebook uses the Smith-Waterman algorithm to find *local* regions of alignment between two pieces of text.

In [ ]:
```
import numpy as np
import re
from nltk import word_tokenize
from nltk.stem import WordNetLemmatizer
```

In [ ]:
```
def base_similarity(token1, token2):
    if token1 == token2:
        return 3
    else:
        return -3
```

In [ ]:
```
def smith_waterman(one, two, similarity_function):
    indel=-2

    # columns
    m=len(one)
    # rows
    n=len(two)
    scores=np.zeros((n+1, m+1))
    backpointers=np.ones((n+1,m+1), dtype=int)*-1

    maxtotal=0
    maxrow=-1
    maxcol=-1
    # rows
    for i in range(1,n+1):
        # columns
        for j in range(1,m+1):
            left=scores[i,j-1]+indel
            top=scores[i-1,j]+indel
            diag=scores[i-1,j-1] + similarity_function(one[j-1], two[i-1])

            maxx=top
            backpointers[i,j]=0
            if left > maxx:
                maxx=left
                backpointers[i,j]=1
            if diag > maxx:
                maxx=diag
                backpointers[i,j]=2

            scores[i,j]=maxx

            if scores[i,j] < 0:
                scores[i,j]=0
```

```
            if scores[i,j] > maxtotal:
                maxtotal=scores[i,j]
                maxrow=i
                maxcol=j


    argscores=np.dstack(np.unravel_index(np.argsort(-scores.ravel()), (n+1, m

    all_alignments=[]
    seen_best={}

    # only show sequences that have a score of 10 or higher
    # (with a similarity_function score of 3, this effectively means >3 words

    minScore=10

    # this loop finds all alignments between the source and target, but only
    # in the source and target sequence to belong to *one* alignment.

    for top_n in range(len(argscores)):

        overlapFlag=False
        source_alignments=[]
        row,col=argscores[top_n]
        score=scores[row,col]
        if score < minScore:
            break

        start=backpointers[row,col]

        this_seen_best={}

        if ("C", col-1) in seen_best or ("R", row-1) in seen_best:
            continue

        while score > 0:
            if start == 0:
                row-=1
            if start == 1:
                col-=1
            if start == 2:
                if one[col-1] == two[row-1] and re.search("\S", one[col-1]) !
                    if ("C", col-1) not in seen_best and ("R", row-1) not in
                        source_alignments.append((col-1, row-1))
                    else:
                        overlapFlag=True
                        break
                    this_seen_best[("C", col-1)]=1
                    this_seen_best[("R", row-1)]=1

                row-=1
                col-=1
```

```
            start=backpointers[row,col]
            score=scores[row,col]

        if not overlapFlag:

            for key in this_seen_best:
                seen_best[key]=1
            maxLeftSource=m
            maxRightSource=0
            maxLeftTarget=n
            maxRightTarget=0
            for s, t in source_alignments:
                if s < maxLeftSource:
                    maxLeftSource=s
                if s > maxRightSource:
                    maxRightSource=s
                if t < maxLeftTarget:
                    maxLeftTarget=t
                if t > maxRightTarget:
                    maxRightTarget=t

            row,col=argscores[top_n]
            all_alignments.append((scores[row,col], source_alignments, maxLef

    return all_alignments
```

In [ ]:

```
def sw_compare(oneString, twoString, similarity_function):
    one=word_tokenize(oneString)
    two=word_tokenize(twoString)

    alignments=smith_waterman(one, two, similarity_function)
    for score, source_alignments, leftSource, rightSource, leftTarget, rightT
        print("Score: %d\n" % score)
        print("one (%d, %d):" % (leftSource, rightSource), ' '.join(one[leftS
        print()
        print("two (%d, %d):" % (leftTarget, rightTarget), ' '.join(two[leftT
        print("\n==========\n")
```

In [ ]:

```
stevie_wonder_pastime_paradise="""Been spending most their lives
Living in a pastime paradise
They've been spending most their lives
Living in a pastime paradise
They've been wasting most their time
Glorifying days long gone behind
They've been wasting most their days
In rememberance of ignorance oldest praise
Tell me who of them will come to be
How many of them are you and me
Dissipation
Race relations
```

```
Consolation
Segregation
Dispensation
Isolation
Exploitation
Mutilation
Mutations
Miscreation
Confirmation to the evils of the world
Been spending most their lives
Living in a future paradise
They've been spending most their lives
Living in a future paradise
They've been looking in their minds
For the days that sorrow's gone from time
They keep telling of the day
When the Savior of love will come to stay
Tell me who of them will come to be
How many of them are you and me
Proclamation
Of race relations
Consolations
Integration
Verification
Of revelations
Acclamation
World salvation
Vibrations
Stimulation
Confirmation to the peace of the world
They've been spending most their lives
Living in a pastime paradise
They've been spending most their lives
Living in a pastime paradise
They've been spending most their lives
Living in a future paradise
They've been spending most their lives
Living in a future paradise
We've been spending too much of our lives
Living in a pastime paradise
Let's start living our lives
Living for the future paradise
Praise to our lives
Living in the future paradise
Shame to anyone lives
Living in a pastime paradise"""
```

In [ ]:

```
coolio_gangtas_paradise="""As I walk through the valley of the shadow of deat
I take a look at my life and realize there's nothing left
'Cause I've been blasting and laughing so long that
Even my momma thinks that my mind is gone
But I ain't never crossed a man that didn't deserve it
```

```
Me be treated like a punk, you know that's unheard of
You better watch how you talking and where you walking
Or you and your homies might be lined in chalk
I really hate to trip, but I gotta loc
As they croak, I see myself in the pistol smoke
Fool, I'm the kinda G the little homies wanna be like
On my knees in the night, saying prayers in the streetlight
Been spending most their lives
Living in a gangsta's paradise
Been spending most their lives
Living in a gangsta's paradise
Keep spending most our lives
Living in a gangsta's paradise
Keep spending most our lives
Living in a gangsta's paradise
Look at the situation they got me facing
I can't live a normal life, I was raised by the street
So I gotta be down with the hood team
Too much television watchin', got me chasing dreams
I'm a educated fool with money on my mind
Got my ten in my hand and a gleam in my eye
I'm a loc'd out gangsta, set tripping banger
And my homies is down, so don't arouse my anger
Fool, death ain't nothing but a heart beat away
I'm living life do or die, what can I say?
I'm 23 now but will I live to see 24?
The way things is going I don't know
Tell me why are we so blind to see
That the ones we hurt are you and me?
Been spending most their lives
Living in a gangsta's paradise
Been spending most their lives
Living in a gangsta's paradise
Keep spending most our lives
Living in a gangsta's paradise
Keep spending most our lives
Living in a gangsta's paradise
Power and the money, money and the power
Minute after minute, hour after hour
Everybody's runnin', but half of them ain't looking
It's going on in the kitchen, but I don't know what's cooking
They say I gotta learn, but nobody's here to teach me
If they can't understand it, how can they reach me?
I guess they can't, I guess they won't
I guess they front, that's why I know my life is out of luck, fool
Been spending most their lives
Living in a gangsta's paradise
Been spending most their lives
Living in a gangsta's paradise
Keep spending most our lives
Living in a gangsta's paradise
Keep spending most our lives
Living in a gangsta's paradise
```

```
Tell me why are we so blind to see
That the ones we hurt are you and me?
Tell me why are we so blind to see
That the ones we hurt are you and me?"""
```

In [ ]:
```
sw_compare(coolio_gangtas_paradise, stevie_wonder_pastime_paradise, base_simi
```

```
Score: 45

one (476, 515): spending most their lives Living in a gangsta 's paradise Been
spending most their lives Living in a gangsta 's paradise Keep spending most o
ur lives Living in a gangsta 's paradise Keep spending most our lives Living i
n a

two (176, 218): spending most their lives Living in a pastime paradise They 'v
e been spending most their lives Living in a pastime paradise They 've been sp
ending most their lives Living in a future paradise They 've been spending mos
t their lives Living in a

==========

Score: 36

one (137, 155): Been spending most their lives Living in a gangsta 's paradise
Been spending most their lives Living in a

two (82, 101): Been spending most their lives Living in a future paradise They
've been spending most their lives Living in a

==========

Score: 36

one (332, 350): Been spending most their lives Living in a gangsta 's paradise
Been spending most their lives Living in a

two (0, 19): Been spending most their lives Living in a pastime paradise They
've been spending most their lives Living in a

==========

Score: 12

one (533, 536): are you and me

two (60, 63): are you and me

==========

Score: 12

one (327, 330): are you and me
```

```
two (148, 151): are you and me

==========

Score: 12

one (369, 372): lives Living in a

two (258, 261): lives Living in a

==========

Score: 12

one (162, 165): our lives Living in

two (248, 251): our lives Living in

==========
```

In [ ]:
```
kjv_bible_proverbs23="""The Lord is my shepherd; I shall not want. He maketh
```

In [ ]:
```
sw_compare(coolio_gangtas_paradise, kjv_bible_proverbs23, base_similarity)
```

```
Score: 31

one (1, 11): I walk through the valley of the shadow of death I

two (52, 63): I walk through the valley of the shadow of death , I

==========
```

In [ ]:
```
weird_al_eat_it="""Just eat it, eat it, eat it, eat it
Get yourself an egg and beat it
Have some more chicken, have some more pie
It doesn't matter if it's broiled or fried
Just eat it, eat it, just eat it, eat it"""
```

In [ ]:
```
jackson_beat_it="""Just beat it, beat it, beat it, beat it
No one wants to be defeated
Showin' how funky and strong is your fight
It doesn't matter who's wrong or right
Just beat it, beat it"""
```

In [ ]:
```
sw_compare(weird_al_eat_it, jackson_beat_it, base_similarity)
```

```
Score: 13

one (28, 43): It does n't matter if it 's broiled or fried Just eat it , eat i
t

two (27, 41): It does n't matter who 's wrong or right Just beat it , beat it

==========

Score: 12

one (2, 11): it , eat it , eat it , eat it

two (2, 11): it , beat it , beat it , beat it

==========
```

In [ ]:
```
wolf_killing_floor="""I should'a quit you, a long time ago
I should'a quit you, baby, long time ago
I should'a quit you, and went on to Mexico

If I ha'da followed my first mind
If I ha'da followed my first mind
I'd'a been gone since my second time"""
```

In [ ]:
```
led_zeppelin_lemon_song="""I should have quit you a long time ago
Ooh - whoa, yeah, yeah, long time ago
I wouldn't be here, my children
Down on this killin' floor
I should have listened, baby, a - to my second mind
Oh, I should have listened, baby, to my second mind"""
```

In [ ]:
```
sw_compare(wolf_killing_floor, led_zeppelin_lemon_song, base_similarity)
```

```
Score: 17

one (0, 10): I should ' a quit you , a long time ago

two (0, 8): I should have quit you a long time ago

==========

Score: 15

one (19, 23): , long time ago I

two (16, 20): , long time ago I

==========
```

Q1. The `base_similarity` method above calculates a very coarse measure of similarity, only testing whether two words are exactly the same. At this point in the course, you have many more methods in your toolbox for thinking about the similarity of two tokens in a sentence. Your only question for this homework is to use that knowledge to develop a better similarity function (`better_similarity`) that captures what you see as the important dimensions of text reuse in these examples.

In [ ]:
```
import nltk
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     /Users/danielfurman/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```
Out[ ]:
```
True
```

In [ ]:
```
def better_similarity(token1, token2):
    lemmatizer = WordNetLemmatizer()
    for i in lemmatizer.lemmatize(token1):
        for j in lemmatizer.lemmatize(token2):
            if i == j:
                return 3
            else:
                return -3
```

In [ ]:
```
sw_compare(coolio_gangtas_paradise, stevie_wonder_pastime_paradise, better_si
```

```
Score: 48

one (138, 180): spending most their lives Living in a gangsta 's paradise Been
spending most their lives Living in a gangsta 's paradise Keep spending most o
```

ur lives Living in a gangsta 's paradise Keep spending most our lives Living i
n a gangsta 's paradise

two (188, 234): spending most their lives Living in a pastime paradise They 'v
e been spending most their lives Living in a future paradise They 've been spe
nding most their lives Living in a future paradise We 've been spending too mu
ch of our lives Living in a pastime paradise

==========

Score: 36

one (475, 493): Been spending most their lives Living in a gangsta 's paradise
Been spending most their lives Living in a

two (0, 19): Been spending most their lives Living in a pastime paradise They
've been spending most their lives Living in a

==========

Score: 36

one (332, 350): Been spending most their lives Living in a gangsta 's paradise
Been spending most their lives Living in a

two (82, 101): Been spending most their lives Living in a future paradise They
've been spending most their lives Living in a

==========

Score: 16

one (509, 518): spending most our lives Living in a gangsta 's paradise

two (176, 184): spending most their lives Living in a pastime paradise

==========

Score: 12

one (533, 536): are you and me

two (148, 151): are you and me

==========

Score: 12

one (519, 520): Tell me

two (47, 48): Tell me

==========

```
Score: 12

one (500, 503): our lives Living in

two (248, 251): our lives Living in

==========

Score: 12

one (327, 330): are you and me

two (60, 63): are you and me

==========

Score: 12

one (358, 361): lives Living in a

two (258, 261): lives Living in a

==========
```

In [ ]:

```
sw_compare(coolio_gangtas_paradise, kjv_bible_proverbs23, better_similarity)
```

```
Score: 31

one (1, 11): I walk through the valley of the shadow of death I

two (52, 63): I walk through the valley of the shadow of death , I

==========

Score: 12

one (185, 187): they got me

two (80, 82): they comfort me

==========

Score: 12

one (107, 108): in the

two (90, 91): in the

==========

Score: 11

one (492, 492): in

two (17, 17): in

==========

Score: 10

one (286, 286): ,

two (62, 62): ,

==========
```

In [ ]:

```
sw_compare(weird_al_eat_it, jackson_beat_it, better_similarity)
```

```
Score: 13

one (28, 41): It does n't matter if it 's broiled or fried Just eat it ,

two (27, 39): It does n't matter who 's wrong or right Just beat it ,

==========

Score: 12

one (2, 9): it , eat it , eat it ,

two (2, 9): it , beat it , beat it ,

==========
```

In [ ]:

```
sw_compare(wolf_killing_floor, led_zeppelin_lemon_song, better_similarity)
```

```
Score: 21

one (0, 23): I should ' a quit you , a long time ago I should ' a quit you , b
aby , long time ago I

two (0, 20): I should have quit you a long time ago Ooh - whoa , yeah , yeah ,
long time ago I

==========
```