NATIONAL UNIVERSITY OF MODERN LANGUAGES

ISLAMABAD

Natural Language Processing (Lab)

Assignment No: 01

Submitted to
Dr. Qurat-ul-Ain Safdar

Submitted By
Junaid Asif
(BSAI-144)

Submission Date: March 18, 2025

1. Extract the English text from the following sentences
   a. This is نیچرل language processing تجربہ گاہ
   b. 你好, I am fine, 你 那

```python
import re

def extract_english(text):
    # Regular expression to match English words
    english_words = re.findall(r'[A-Za-z]+', text)
    return ' '.join(english_words)

# Example sentences
text1 = "This is لرچین language processing ہبرجت ہاگ"
text2 = "你好, I am fine, 你 那"

# Extract English text
print(extract_english(text1))
print(extract_english(text2))
```

✓ 0.0s

```
This is language processing
I am fine
```

2. Explore the library strop words for different languages (e.g. Chinese, Arabic, Japanese, etc.)

```python
import nltk
from nltk.corpus import stopwords


# Available languages
print(stopwords.fileids())
```

✓ 52.1s

['albanian', 'arabic', 'azerbaijani', 'basque', 'belarusian', 'bengali', 'catalan', 'chinese', 'danish', 'dutch', 'english', 'finnish', 'french', 'german', 'greek', 'hebrew', 'hinglish', 'hungarian', 'indonesian', 'italian', 'kazakh', 'nepali', 'norwegian', 'portuguese', 'romanian', 'russian', 'slovene', 'spanish', 'swedish', 'tajik', 'tamil', 'turkish']

```
print("*****Chinese Stopwords*****")
# Load Chinese stopwords
print(stopwords.words('chinese'))
```
✓ 0.0s

*******Chinese Stopwords*******

['一', '一下', '一些', '一切', '一则', '一天', '一定', '一方面', '一旦', '一时', '一来', '一样', '一次', '一片', '一直', '一致', '一般', '一起', '一边', '一面', '万一', '上下', '上升', '上去', '上来', '上述', '上面', '下列', '下去', '下来', '下面', '不一', '不久', '不仅', '不会', '不但', '不光', '不单', '不变', '不只', '不可', '不同', '不够', '不如', '不得', '不怕', '不惟', '不成', '不拘', '不敢', '不断', '不是', '不比', '不然', '不特', '不独', '不管', '不能', '不要', '不论', '不足', '不过', '不问', '与', '与其', '与否', '与此同时', '专门', '且', '两者', '严格', '严重', '个', '个人', '个别', '中小', '中间', '丰富', '临', '为', '为主', '为了', '为什么', '为什麽', '为何', '为着', '主张', '主要', '举行', '乃', '乃至', '么', '之', '之一', '之前', '之后', '之後', '之所以', '之类', '乌乎', '乎', '乘', '也', '也好', '也是', '也罢', '了', '了解', '争取', '于', '于是', '于是乎', '云云', '互相', '产生', '人们', '人家', '什么', '什么样', '什麽', '今后', '今天', '今年', '今後', '仍然', '从', '从事', '从而', '他', '他人', '他们', '他的', '代替', '以', '以上', '以下', '以为', '以便', '以免', '以前', '以及', '以后', '以外', '以後', '以来', '以至', '以至于', '以致', '们', '任', '任何', '任凭', '任务', '企图', '伟大', '似乎', '似的', '但', '但是', '何', '何况', '何处', '何时', '作为', '你', '你们', '你的', '使得', '使用', '例如', '依', '依照', '依靠', '促进', '保持', '俺', '俺们', '倘', '倘使', '倘或', '倘然', '倘若', '假使', '假如', '假若', '做到', '像', '允许', '充分', '先后', '先後', '先生', '全部', '全面', '兮', '共同', '关于', '其', '其一', '其中', '其二', '其他', '其余', '其它', '其实', '其次', '具体', '具体地说', '具体说来', '具有', '再者', '再说', '冒', '冲', '决定', '况且', '准备', '几', '几乎', '几时', '凭', '凭借', '出去', '出来', '出现', '分别', '则', '别', '别的', '别说', '到', '前后', '前者', '前进', '前面', '加之', '加以', '加入', '加强', '十分', '即', '即令', '即使', '即便', '即或', '即若', '却不', '原来', '又', '及', '及其', '及时', '及至', '双方', '反之', '反应', '反映', '反过来', '反过来说', '取得', '受到', '变成', '另', '另一方面', '另外', '只是', '只有', '只要', '只限', '叫', '叫做', '召开', '叮咚', '可', '可以', '可是', '可能', '可见', '各', '各个', '各人', '各位', '各地', '各种', '各级', '各自', '合理', '同', '同一', '同时', '同样', '后来', '后面', '向', '向着', '吓', '吗', '否则', '吧', '吧哒', '吱', '呀', '呃', '呕', '呗', '呜', '呜呼', '呢', '周围', '呵', '呸', '呼哧', '咋', '和', '咚', '咦', '咱', '咱们', '咳', '哇', '哈', '哈哈', '哉', '哎', '哎呀', '哎哟', '哗', '哟', '哦', '哩', '哪', '哪个', '哪些', '哪儿', '哪天', '哪年', '哪怕', '哪样', '哪边', '哪里', '哼', '哼唷', '唉', '啊', '啐', '啥', '啦', '啪达', '喂', '喏', '喔唷', '喽喽', '嗡', '嗯', '嗳', '嘎', '嘎登', '嘘', '嘛', '嘻', '嘿', '因', '因为', '因此', '因而', '固然', '在', '在下', '地', '坚决', '坚持', '基本', '处理', '复杂', '多', '多少', '多

数', '多次', '大力', '大多数', '大大', '大家', '大批', '大约', '大量', '失去', '她', '她们', '她的', '好的', '好象', '如', '如上所述', '如下', '如何', '如其', '如果', '如此', '如若', '存在', '宁', '宁可', '宁愿', '宁肯', '它', '它们', '它们的', '它的', '安全', '完全', '完成', '实现', '实际', '宣布', '容易', '密切', '对', '对于', '对应', '将', '少数', '尔后', '尚且', '尤其', '就', '就是', '就是说', '尽', '尽管', '属于', '岂但', '左右', '巨大', '巩固', '己', '已经', '帮助', '常常', '并', '并不', '并不是', '并且', '并没有', '广大', '广泛', '应当', '应用', '应该', '开外', '开始', '开展', '引起', '强烈', '强调', '归', '当', '当前', '当时', '当然', '当着', '形成', '彻底', '彼', '彼此', '往', '往往', '待', '後来', '後面', '得', '得出', '得到', '心里', '必然', '必要', '必须', '怎', '怎么', '怎么办', '怎么样', '怎样', '怎麽', '总之', '总是', '总的来看', '总的来说', '总的说来', '总结', '总而言之', '恰恰相反', '您', '意思', '愿意', '慢说', '成为', '我', '我们', '我的', '或', '或是', '或者', '战斗', '所', '所以', '所有', '所谓', '打', '扩大', '把', '抑或', '拿', '按', '按照', '换句话说', '换言之', '据', '掌握', '接着', '接著', '故', '故此', '整个', '方便', '方面', '旁人', '无宁', '无法', '无论', '既', '既是', '既然', '时候', '明显', '明确', '是', '是否', '是的', '显然', '显著', '普通', '普遍', '更加', '曾经', '替', '最后', '最大', '最好', '最後', '最近', '最高', '有', '有些', '有关', '有利', '有力', '有所', '有效', '有时', '有点', '有的', '有着', '有著', '望', '朝', '朝着', '本', '本着', '来', '来着', '极了', '构成', '果然', '果真', '某', '某个', '某些', '根据', '根本', '欢迎', '正在', '正如', '正常', '此', '此外', '此时', '此间', '毋宁', '每', '每个', '每天', '每年', '每当', '比', '比如', '比方', '比较', '毫不', '没有', '沿', '沿着', '注意', '深入', '清楚', '满足', '漫说', '焉', '然则', '然后', '然後', '然而', '照', '照着', '特别是', '特殊', '特点', '现代', '现在', '甚么', '甚而', '甚至', '用', '由', '由于', '由此可见', '的', '的话', '目前', '直到', '直接', '相似', '相信', '相反', '相同', '相对', '相对而言', '相应', '相当', '相等', '省得', '看出', '看到', '看来', '看看', '看见', '真是', '真正', '着', '着呢', '矣', '知道', '确定', '离', '积极', '移动', '突出', '突然', '立即', '第', '等', '等等', '管', '紧接着', '纵', '纵令', '纵使', '纵然', '练习', '组成', '经', '经常', '经过', '结合', '结果', '给', '绝对', '继续', '继而', '维持', '综上所述', '罢了', '考虑', '者', '而', '而且', '而况', '而外', '而已', '而是', '而言', '联系', '能', '能否', '能够', '腾', '自', '自个儿', '自从', '自各儿', '自家', '自己', '自身', '至', '至于', '良好', '若', '若是', '若非', '范围', '莫若', '获得', '虽', '虽则', '虽然', '虽说', '行为', '行动', '表明', '表示', '被', '要', '要不', '要不是', '要不然', '要么', '要是', '要求', '规定', '觉得', '认为', '认真', '认识', '让', '许多', '论', '设使', '设若', '该', '说明', '诸位', '谁', '谁知', '赶', '起', '起来', '起见', '趁', '趁着', '越是', '跟', '转动', '转变', '转贴', '较', '较之', '边', '达到', '迅速', '过', '过去', '过来', '运用', '还是', '还有', '这', '这个', '这么', '这么些', '这么样', '这么点儿', '这些', '这会儿', '这儿', '这就是说', '这时', '这样', '这点', '这种', '这边', '这里', '这麽', '进入', '进步', '进而', '进行', '连', '连同', '适应', '适当', '适用', '逐步', '逐渐', '通常', '通过', '造成', '遇到', '遭到', '避免', '那', '那个', '那么', '那么些', '那么样', '那些', '那会儿', '那

儿', '那时', '那样', '那边', '那里', '那麼', '部分', '鄙人', '采取', '里面', '重大', '重新', '重要', '鉴于', '问题', '防止', '阿', '附近', '限制', '除', '除了', '除此之外', '除非', '随', '随着', '随著', '集中', '需要', '非但', '非常', '非徒', '靠', '顺', '顺着', '首先', '高兴', '是不是']

```
print("*****Arabic Stopwords*****")
# Load Arabic stopwords
print(stopwords.words('arabic'))
```
✓ 0.0s

*******Arabic Stopwords*******

['إذ', 'إذا', 'إذما', 'إذن', 'أف', 'أقل', 'أكثر', 'ألا', 'إلا', 'التي', 'الذي', 'الذين', 'اللاتي', 'اللائي', 'اللتان', 'اللتيا', 'اللتين', 'اللذان', 'اللذين', 'اللواتي', 'إلى', 'إليك', 'إليكم', 'إليكما', 'إليكن', 'أم', 'أما', 'أما', 'إما', 'أن', 'إن', 'إنا', 'أنا', 'أنت', 'أنتم', 'أنتما', 'أنتن', 'إنما', 'إنه', 'أنى', 'أنى', 'آه', 'آها', 'أو', 'أولاء', 'أولئك', 'أوه', 'آي', 'أي', 'أيها', 'إي', 'أين', 'أين', 'أينما', 'إيه', 'بخ', 'بس', 'بعد', 'بعض', 'بك', 'بكم', 'بكم', 'بكما', 'بكن', 'بل', 'بلى', 'بما', 'بماذا', 'بمن', 'بنا', 'به', 'بها', 'بهم', 'بهما', 'بهن', 'بي', 'بين', 'بيد', 'تلك', 'تلكم', 'تلكما', 'ته', 'تي', 'تين', 'تينك', 'ثم', 'ثمة', 'حاشا', 'حبذا', 'حتى', 'حيث', 'حيثما', 'حين', 'خلا', 'دون', 'ذا', 'ذات', 'ذاك', 'ذان', 'ذانك', 'ذلك', 'ذلكم', 'ذلكما', 'ذلكن', 'ذه', 'ذو', 'ذوا', 'ذواتا', 'ذواتي', 'ذي', 'ذين', 'ذينك', 'ريث', 'سوف', 'سوى', 'شتان', 'عدا', 'عسى', 'عل', 'على', 'عليك', 'عليه', 'عما', 'عن', 'عند', 'غير', 'فإذا', 'فإن', 'فلا', 'فمن', 'في', 'فيم', 'فيما', 'فيه', 'فيها', 'قد', 'كأن', 'كأنما', 'كأي', 'كأين', 'كذا', 'كذلك', 'كل', 'كلا', 'كلاهما', 'كلتا', 'كلما', 'كليكما', 'كليهما', 'كم', 'كم', 'كما', 'كي', 'كيت', 'كيف', 'كيفما', 'لا', 'لاسيما', 'لدى', 'لست', 'لستم', 'لستما', 'لستن', 'لسن', 'لسنا', 'لعل', 'لك', 'لكم', 'لكما', 'لكن', 'لكنما', 'لكي', 'لكيلا', 'لم', 'لما', 'لن', 'لنا', 'له', 'لها', 'لهم', 'لهما', 'لهن', 'لو', 'لولا', 'لوما', 'لي', 'لئن', 'ليت', 'ليس', 'ليسا', 'ليست', 'ليستا', 'ليسوا', 'ما', 'ماذا', 'متى', 'مذ', 'مع', 'مما', 'ممن', 'من', 'منه', 'منها', 'منذ', 'مه', 'مهما', 'نحن', 'نحو', 'نعم', 'ها', 'هاتان', 'هاته', 'هاتي', 'هاتين', 'هاك', 'هاهنا', 'هذا', 'هذان', 'هذه', 'هذي', 'هذين', 'هكذا', 'هل', 'هلا', 'هم', 'هما', 'هن', 'هنا', 'هناك', 'هنالك', 'هو', 'هؤلاء', 'هي', 'هيا', 'هيت', 'هيهات', 'والذي', 'والذين', 'وإذ', 'وإذا', 'وإن', 'ولا', 'ولكن', 'ولو', 'وما', 'ومن', 'وهو', 'يا', 'أبٌ', 'أخٌ', 'حمٌ', 'فو', 'أنتِ', 'يناير', 'فبراير', 'مارس', 'أبريل', 'مايو', 'يونيو', 'يوليو', 'أغسطس', 'سبتمبر', 'أكتوبر', 'نوفمبر', 'ديسمبر', 'جانفي', 'فيفري', 'مارس', 'أفريل', 'ماي', 'جوان', 'جويلية', 'أوت', 'كانون', 'شباط', 'آذار', 'نيسان', 'أيار', 'حزيران', 'تموز', 'آب', 'أيلول', 'تشرين', 'دولار', 'دينار', 'ريال', 'درهم', 'ليرة', 'جنيه', 'قرش', 'مليم', 'فلس', 'هللة', 'سنتيم', 'يورو', 'ين', 'يوان', 'شيكل', 'واحد', 'اثنان', 'ثلاثة', 'أربعة', 'خمسة', 'ستة', 'سبعة', 'ثمانية', 'تسعة', 'عشرة', 'أحد', 'اثنا', 'اثني', 'إحدى', 'ثلاث', 'أربع', 'خمس', 'ست', 'سبع', 'ثماني', 'تسع', 'عشر', 'ثمان', 'سبت', 'أحد', 'اثنين', 'ثلاثاء', 'أربعاء', 'خميس', 'جمعة', 'أول', 'ثان', 'ثاني', 'ثالث', 'رابع', 'خامس', 'سادس', 'سابع', 'ثامن', 'تاسع', 'عاشر', 'حادي', 'أ', 'ب', 'ت', 'ث', 'ج', 'ح', 'خ', 'د', 'ذ', 'ر', 'ز', 'س', 'ش', 'ص', 'ض', 'ط', 'ظ', 'ع', 'غ', 'ف', 'ق', 'ك', 'ل', 'م', 'ن', 'ه', 'و', 'ي', 'ء', 'ى', 'آ', 'ؤ', 'ئ', 'أ', 'ة', 'ألف', 'باء', 'تاء', 'ثاء', 'جيم', 'حاء', 'خاء', 'دال', 'ذال', 'راء', 'زاي', 'سين', 'شين', 'صاد', 'ضاد', 'طاء', 'ظاء', 'عين', 'غين', 'فاء', 'قاف', 'كاف', 'لام', 'ميم', 'نون', 'هاء', 'واو', 'ياء', 'همزة', 'ي', 'نا', 'ك', 'كن', 'ه', 'إياه', 'إياها', 'إياهما', 'إياهم', 'إياهن', 'إياك', 'إياكما', 'إياكم', 'إياك', 'إياكن', 'إياي', 'إيانا', 'أولالك', 'تانِ', 'تانِك', 'تِه', 'تِي', 'تَيْنِ', 'ثمّ', 'ثمّة', 'ذانِ', 'ذِه', 'ذِي', 'ذَيْنِ', 'هَؤلاء', 'هَاتانِ', 'هَاتِه', 'هَاتِي', 'هَاتَيْنِ', 'هَذا', 'هَذانِ', 'هَذِه', 'هَذِي', 'هَذَيْنِ', 'الأُلى', 'الأولى', 'الأُلاء', 'أل', 'أنّى', 'أيّ', 'أيّان', 'أنّى', 'أيّ', 'أيّان', 'ذيت', 'كأيّ', 'كأيّن', 'بضع', 'فلان', 'وا', 'آمينَ', 'آهِ', 'آهٍ', 'آهاً', 'أُفٍّ', 'أُفٍّ', 'أفٍّ', 'أمامك', 'أمامكَ', 'أوّهْ', 'إلَيْكَ', 'إلَيْكَ', 'إليكَ', 'إليكنّ', 'إيهٍ', 'بخٍ', 'بسّ', 'بَسْ', 'بطآن', 'بَلْهَ', 'حاي', 'حَذارِ', 'حيَّ', 'حيَّ', 'دونك', 'رويدك', 'سرعان', 'شتانَ', 'شَتَّانَ', 'صهْ', 'صهٍ', 'طاق', 'طَق', 'عَدَسْ', 'كِخ', 'مكانَك', 'مكانَك', 'مكانَك', 'مكانكم', 'مكانكما', 'مكانكنّ', 'نَخْ', 'هاكَ', 'هَجْ', 'هلم', 'هيّا', 'هَيْهات', 'وا', 'واهاً', 'وراءَك', 'وُشْكَانَ', 'وَيْ', 'يفعلان', 'تفعلان', 'يفعلون', 'تفعلون', 'تفعلين', 'اتخذ', 'ألفى', 'تخذ', 'ترك', 'تعلَّم', 'جعل', 'حجا', 'حبيب', 'خال', 'حسب', 'خال', 'درى', 'رأى', 'زعم', 'صبر', 'ظنَّ', 'عدَّ', 'علم', 'غادر', 'ذهب', 'وجد', 'ورد', 'وهب', 'أسكن', 'أطعم', 'أعطى', 'رزق', 'زود', 'سقى', 'كسا', 'أخبر', 'أرى', 'أعلم', 'أنبأ', 'حدَث', 'خبَّر', 'نبَّا', 'أفعل به', 'ما أفعله', 'بئس', 'ساء', 'طالما', 'قلما', 'لات', 'لكنَّ', 'ءَ', 'أجل', 'إذاً', 'أمّا', 'إمّا', 'إنَّ', 'أنَّ', 'أى', 'إى', 'أيا', 'ب', 'ثمَّ', 'جلل', 'جير', 'رُبَّ', 'س', 'علَّ', 'ف', 'كأنّ', 'كلَّا', 'كى', 'ل',

'الات', 'لعلَّ', 'لكنَّ', 'لكنَّ', 'م', 'نَّ', 'هلّا', 'وا', 'أل', 'إلّا', 'ت', 'ك', 'لمّا', 'ن', 'ه', 'و', 'ا', 'ي', 'تجاه', 'تلقاء', 'جميع', 'حسب', 'سبحان', 'شبه', 'لعمر', 'مثل', 'معاذ', 'أبو', 'أخو', 'حمو', 'فو', 'مئة', 'مئتان', 'ثلاثمئة', 'أربعمئة', 'خمسمئة', 'ستمئة', 'سبعمئة', 'ثمنمئة', 'تسعمئة', 'مائة', 'ثلاثمائة', 'أربعمائة', 'خمسمائة', 'ستمائة', 'سبعمائة', 'ثمانمئة', 'تسعمائة', 'عشرون', 'ثلاثون', 'اربعون', 'خمسون', 'ستون', 'سبعون', 'ثمانون', 'تسعون', 'عشرين', 'ثلاثين', 'اربعين', 'خمسين', 'ستين', 'سبعين', 'ثمانين', 'تسعين', 'بضع', 'نيف', 'أجمع', 'جميع', 'عامة', 'عين', 'نفس', 'لا سيما', 'أصلا', 'أهلا', 'أيضا', 'بؤسا', 'بعدا', 'بغتة', 'تعسا', 'حقا', 'حمدا', 'خلافا', 'خاصة', 'دواليك', 'سحقا', 'سرا', 'سمعا', 'صبرا', 'صدقا', 'صراحة', 'طرا', 'عجبا', 'عيانا', 'غالبا', 'فرادى', 'فضلا', 'قاطبة', 'كثيرا', 'لبيك', 'معاذ', 'أبدا', 'إزاء', 'أصلا', 'الآن', 'أمد', 'أمس', 'آنفا', 'آناء', 'أنّى', 'أول', 'أيّان', 'تارة', 'ثمّ', 'ثمّة', 'حقا', 'صباح', 'مساء', 'ضحوة', 'عوض', 'غدا', 'غداة', 'قطّ', 'كلّما', 'لدن', 'لمّا', 'مرّة', 'قبل', 'خلف', 'أمام', 'فوق', 'تحت', 'يمين', 'شمال', 'ارتدّ', 'استحال', 'أصبح', 'أضحى', 'آض', 'أمسى', 'انقلب', 'بات', 'تبدّل', 'تحوّل', 'حار', 'رجع', 'راح', 'صار', 'ظلّ', 'عاد', 'غدا', 'كان', 'ما انفك', 'ما برح', 'مادام', 'مازال', 'مافتئ', "ابتدأ", 'أخذ', 'اخلولق', 'أقبل', 'انبرى', 'أنشأ', 'أوشك', 'جعل', 'حرى', 'شرع', 'طفق', 'علق', 'قام', 'كرب', 'كاد', 'هبّ']

Japanese Stop Words:

Unfortunately, NLTK does not provide built-in Japanese stop words like it does for English (stopwords.words('english')). However, you can manually add Japanese stop words to NLTK's stopword list.

3. Differentiate between lemmatization and stemming with the help of exemplary code.

Feature	Lemmatization	Stemming
Definition	Converts word to its dictionary root (lemma).	Removes word suffixes to get the base form.
Accuracy	More accurate, uses word meaning.	Less accurate, may produce non-real words.
Example	running → run, better → good	running → runn, better → better
Uses	NLP models, text processing requiring real words.	Quick text preprocessing where accuracy is not a concern.

```python
import nltk
from nltk.stem import PorterStemmer, WordNetLemmatizer

# Initialize stemmer and lemmatizer
ps = PorterStemmer()
lemmatizer = WordNetLemmatizer()

word = "better"

# Stemming
stemmed_word = ps.stem(word)

# Lemmatization (as a verb)
lemmatized_word = lemmatizer.lemmatize(word, pos='a')

print(f"Original: {word}")
print(f"Stemming: {stemmed_word}")
print(f"Lemmatization: {lemmatized_word}")
```

✓ 0.0s

```
Original: better
Stemming: better
Lemmatization: good
```