**POC: Homgraphy Word Detector  
  
What Is Homography (Homoglyphs)?**

Homoglyphs are letters or symbols from different alphabets that look almost identical but are actually different characters. For example, the Latin letter **“a”**, the Cyrillic letter **“а”**, and the Greek letter **“α”** all look very similar but are different Unicode characters.

Because of this similarity, attackers sometimes use homoglyphs to create fake words or domain names to trick people. For example, the fake domain **раураl.com** (with Cyrillic letters) looks like **paypal.com** but is different and potentially dangerous.

**ASCII vs Unicode**

* **ASCII** is an old character encoding system that represents basic English letters, digits, and symbols. It only supports 128 characters (codes 0–127), covering standard English text.
* **Unicode** is a universal character encoding that supports over a million characters from many languages, scripts, emojis, and symbols. This includes Latin, Greek, Cyrillic alphabets, and more.

Since Unicode includes many alphabets, some characters look very similar visually but have different code points. That’s where homoglyphs come in.

**Similar Characters Across Alphabets**

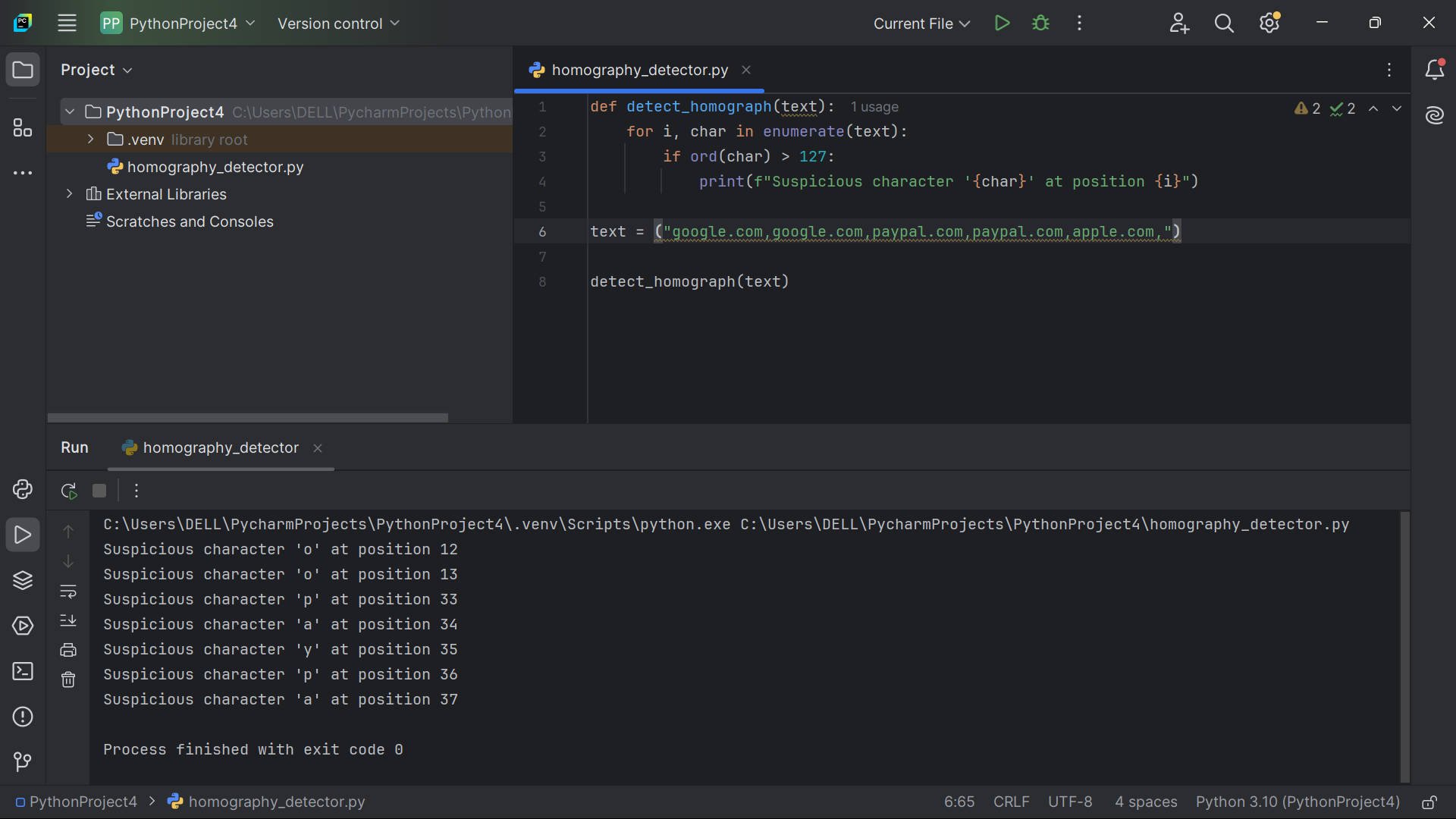
Some letters look the same but belong to different alphabets:

| **Latin** | **Cyrillic** | **Greek** |
| --- | --- | --- |
| A | А (U+0410) | Α (U+0391) |
| a | а (U+0430) | α (U+03B1) |
| e | е (U+0435) | ε (U+03B5) |
| o | о (U+043E) | ο (U+03BF) |
| p | р (U+0440) | ρ (U+03C1) |

Because computers see these as distinct characters, software can detect when someone uses a homoglyph.

**Simple Conceptual Proof of Concept (PoC)**

1. Have a list of words you want to check.
2. For each word, check each character’s Unicode code.
3. If any character has a Unicode value greater than 127, it might be a homoglyph (not plain ASCII).
4. Flag the word as suspicious if it contains such characters.



**Why This Matters**

Homography (Homoglyphs) are letters from different alphabets that look almost the same but are different characters. For example, the Latin letter "a", the Cyrillic "а", and the Greek "α" look very similar but have different codes.

This can be used to trick people by writing fake words or domain names that look real but use these similar-looking letters.

ASCII is a basic character set with 128 characters, mostly English letters and symbols. Unicode is a much larger system that includes letters from many languages, allowing characters that look alike but come from different alphabets.

Because of this, detecting homoglyphs involves checking if letters come from outside the basic ASCII set — those might be fakes trying to confuse you.